

Theoretical Neuroscience and Deep Learning Theory

Surya Ganguli

Dept. of Applied Physics,
Neurobiology,
and Electrical Engineering

Stanford University

Funding:

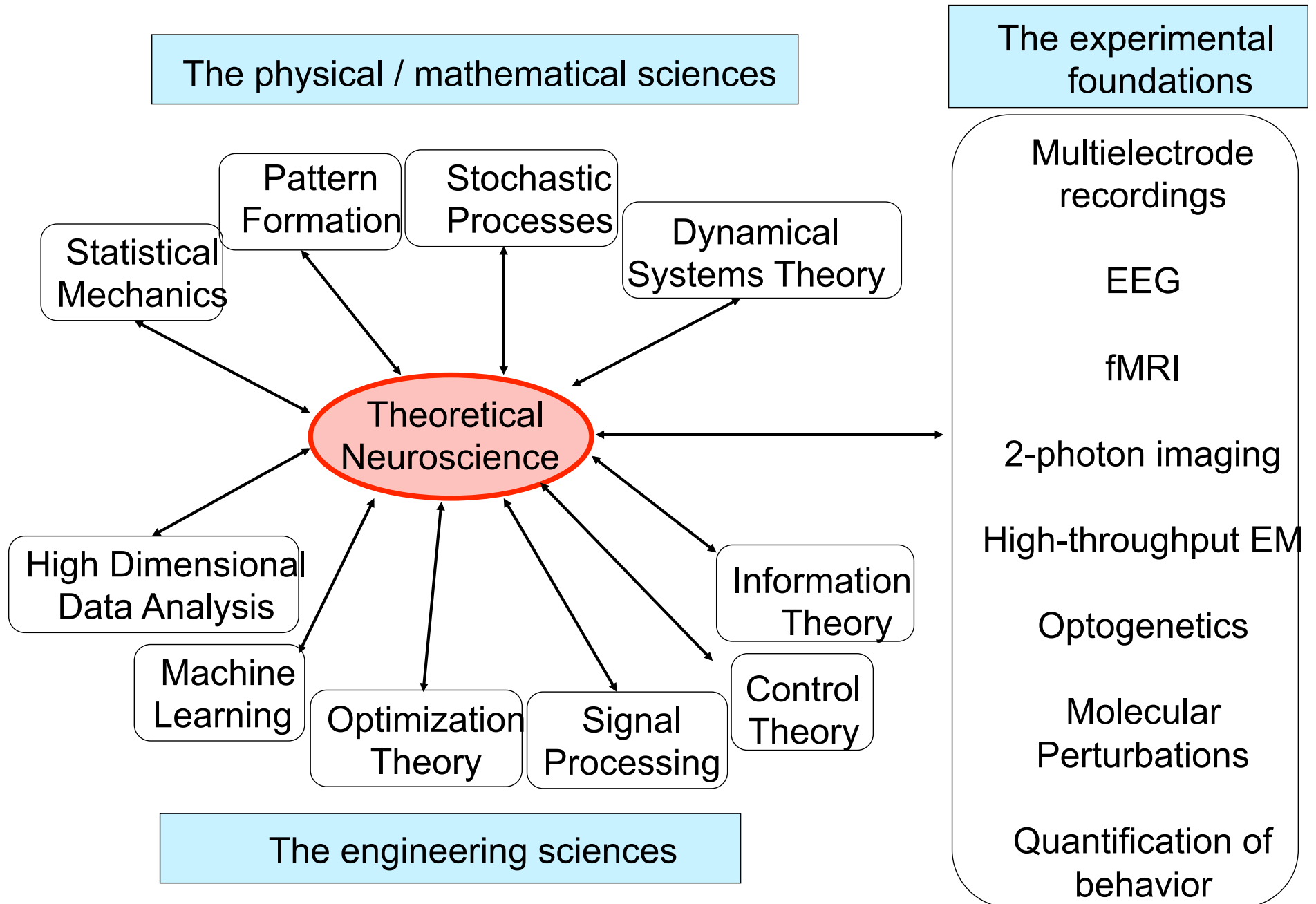
Bio-X Neuroventures
Burroughs Wellcome
Genentech Foundation
James S. McDonnell Foundation
McKnight Foundation
National Science Foundation

NIH
Office of Naval Research
Simons Foundation
Sloan Foundation
Swartz Foundation
Stanford Terman Award

<http://ganguli-gang.stanford.edu>

Twitter: @SuryaGanguli

Theoretical neuroscience in the disciplinary landscape



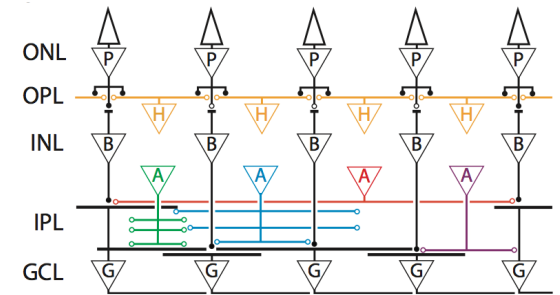
Neural circuits and behavior: theory, computation and experiment

with **Baccus lab**: inferring hidden circuits in the retina

w/ Niru Maheswaranathan and Lane McIntosh

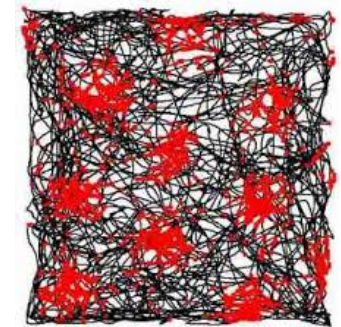
with **Clandinin lab**: unraveling the computations underlying fly motion vision from whole brain optical imaging

w/ Jonathan Leong, Ben Poole and Jennifer Esch



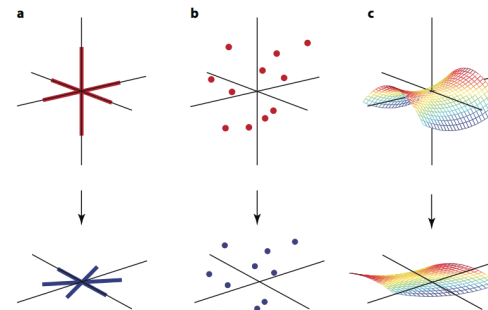
with the **Giocomo lab**: understanding the internal representations of space in the mouse entorhinal cortex

w/ Kiah Hardcastle and Sam Ocko



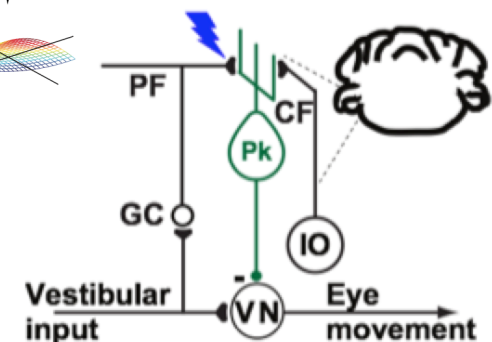
with the **Shenoy lab**: a theory of neural dimensionality, dynamics and measurement

w/ Peiran Gao, Eric Trautmann, and Chris Stock



with the **Raymond lab**: theories of how enhanced plasticity can either enhance or impair learning depending on experience

w/ Subhanil Lahiri, Barbara Vu, Grace Zhao



Motivations for an alliance between **theoretical neuroscience** and theoretical **machine learning**

- What does it mean to understand the brain (or a neural circuit?)
- We understand how the connectivity and dynamics of a neural circuit gives rise to behavior.
- And also how neural activity and synaptic learning rules conspire to self-organize useful connectivity that subserves behavior.
- It is a good start, but it is not enough, to develop a theory of either random networks that have no function.
- The field of machine learning has generated a plethora of learned neural networks that accomplish interesting functions.
- We know their connectivity, dynamics, learning rule, and developmental experience, **yet**, we do not have a meaningful understanding of how they learn and work!

On simplicity and complexity in the brave new world of large scale neuroscience, Peiran Gao and S. Ganguli, Curr. Op. in Neurobiology, 2015.

Talk Outline

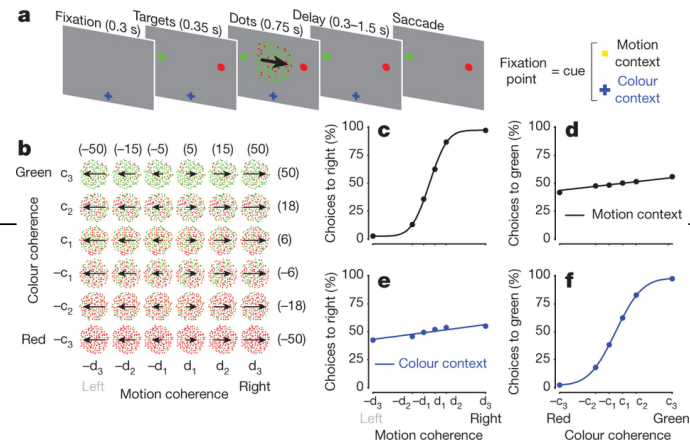
- **Applying deep learning to the brain:**
 - Recurrent neural networks for context dependent decision making
 - Feed-forward networks for modeling the ventral visual stream
 - State of the art models of retinal function
- **Theory of deep learning:**
 - Optimization
 - Expressivity
 - Generalization
- **Inspiration from neuroscience back to deep learning:**
 - Canonical cortical microcircuits
 - Nested loop architectures
 - Avoiding catastrophic forgetting through synaptic complexity
 - Learning asymmetric recurrent generative models

The shape of things to come... on monkeys and models

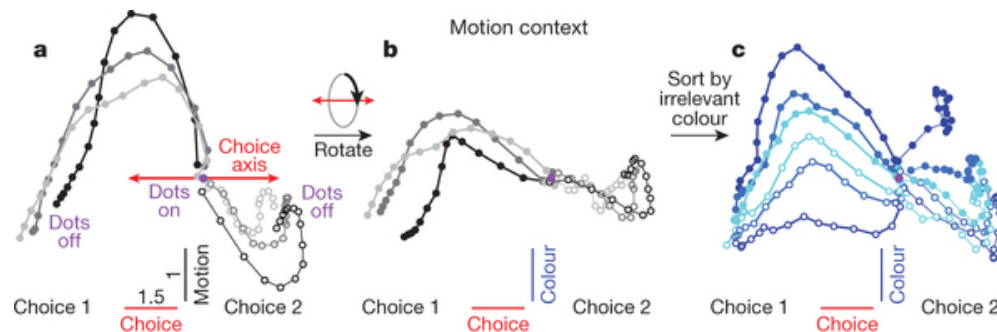
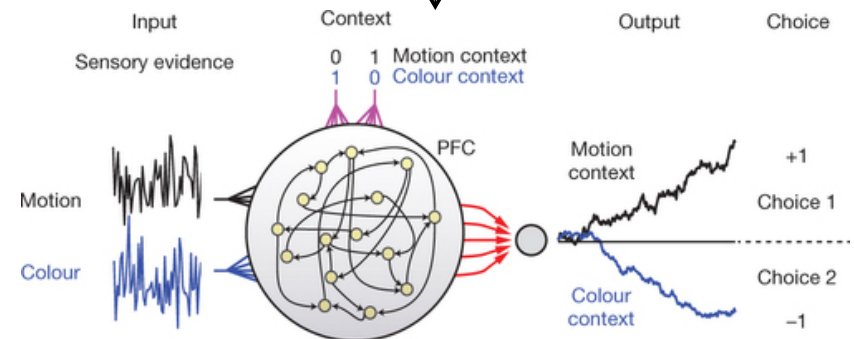
Mante et.al. Context dependent computation by recurrent dynamics in prefrontal cortex, Nature 2013

A behavioral task

The monkey

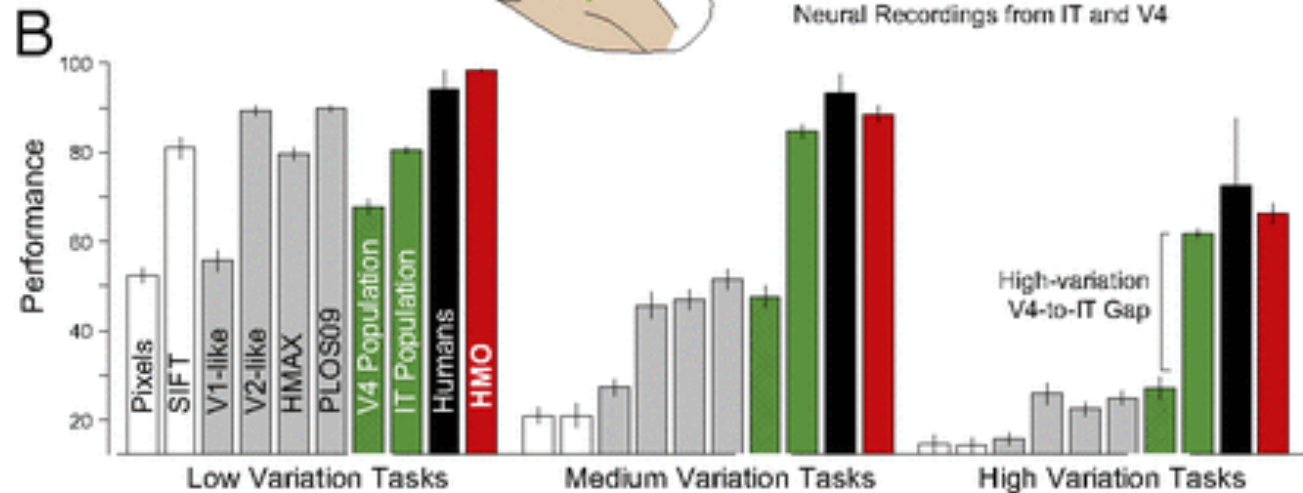
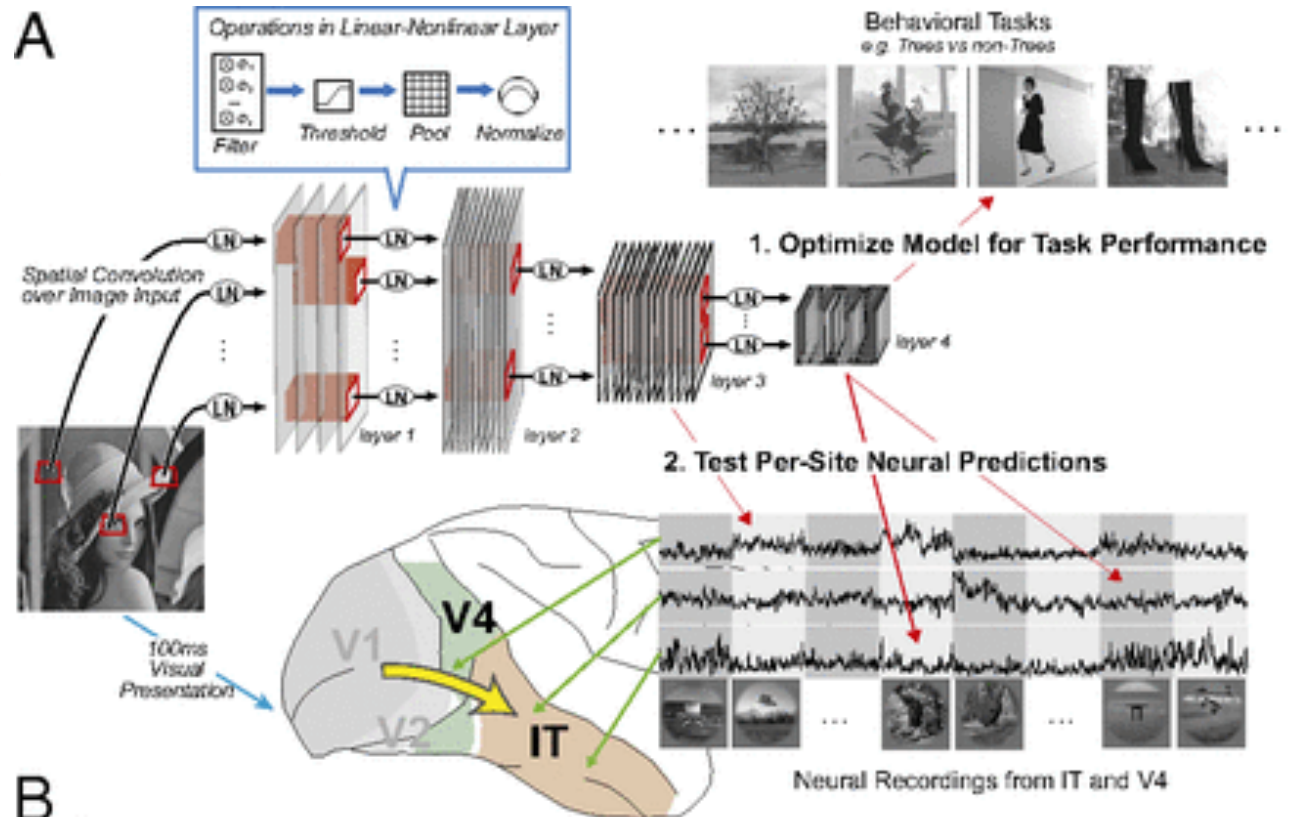


The model

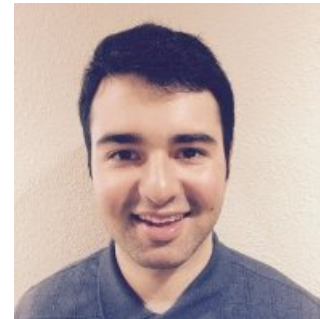


The shape of things to come... on monkeys and models

Yamins et.al. Performance optimized hierarchical models predict performance in higher visual cortex, PNAS 2014



Deep neural network models of the retinal response to natural scenes

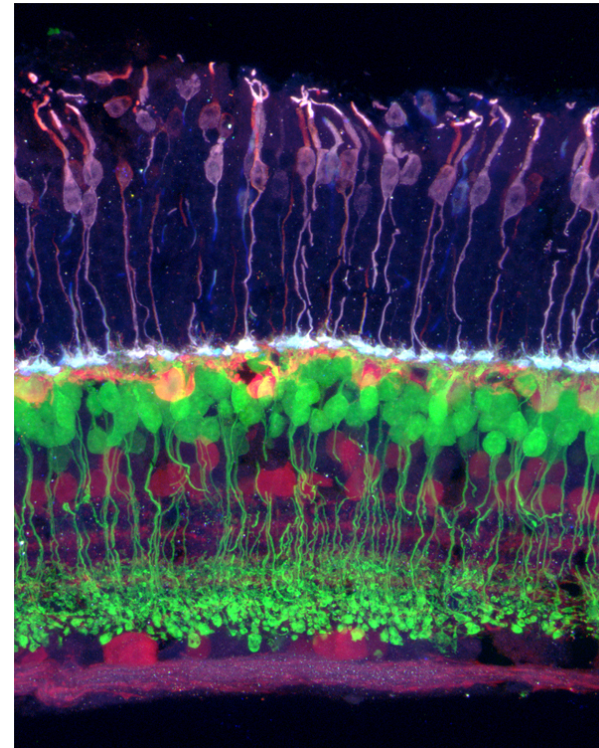
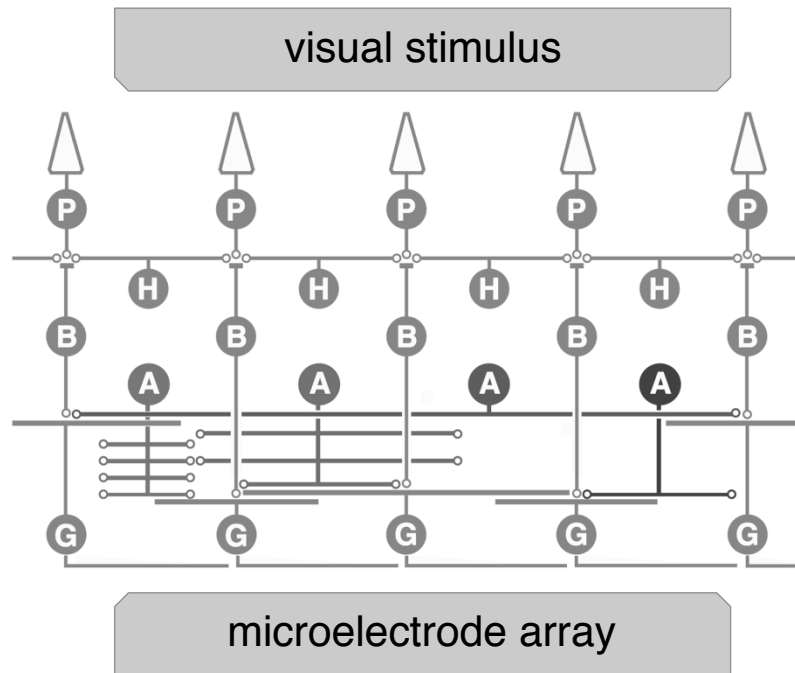


Lane McIntosh and Niru Maheswaranathan, Aran Nayebi,
Surya Ganguli and Stephen Baccus



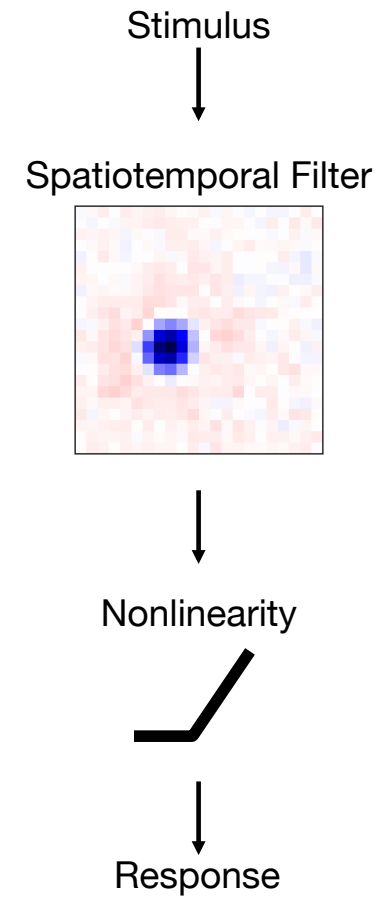
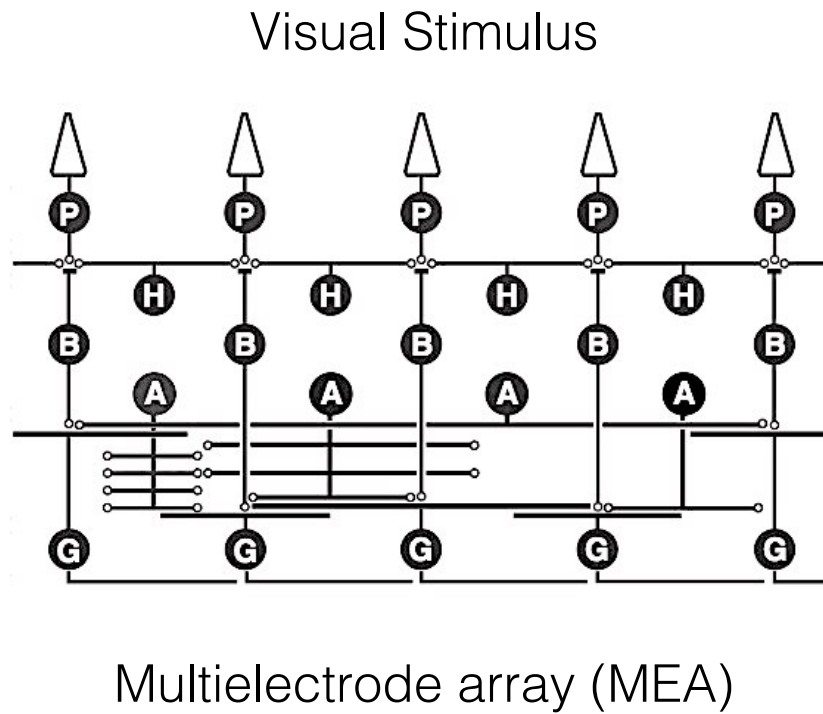
McIntosh, L.*; Maheswaranathan, N.*; Nayebi, A., Ganguli, S.,
Baccus, S.A. *Deep Learning Models of the Retinal Response to
Natural Scenes*. NIPS 2016.

A brief tour of the retina



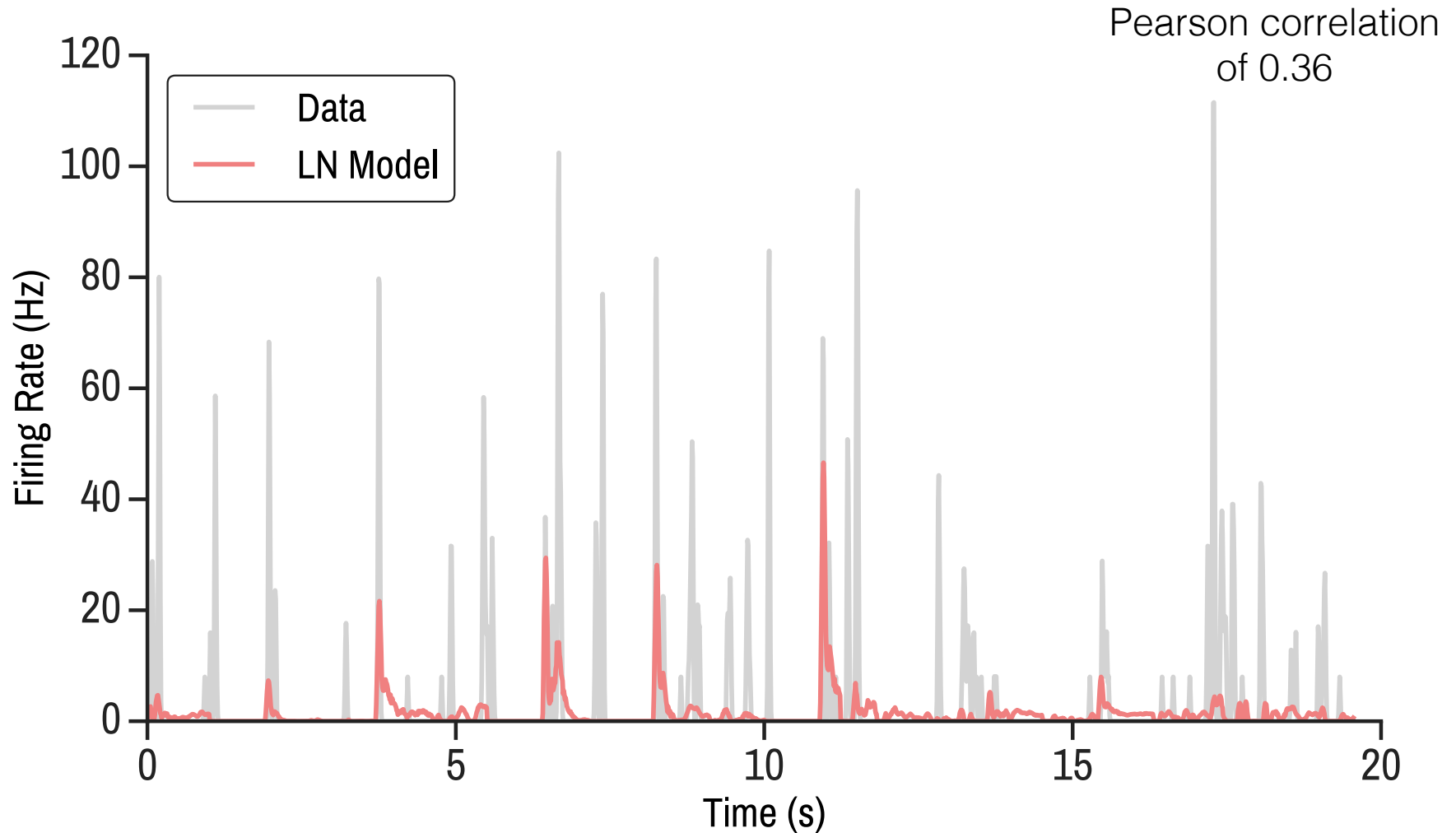
From Rachel Wong's Lab

Linear-Nonlinear models



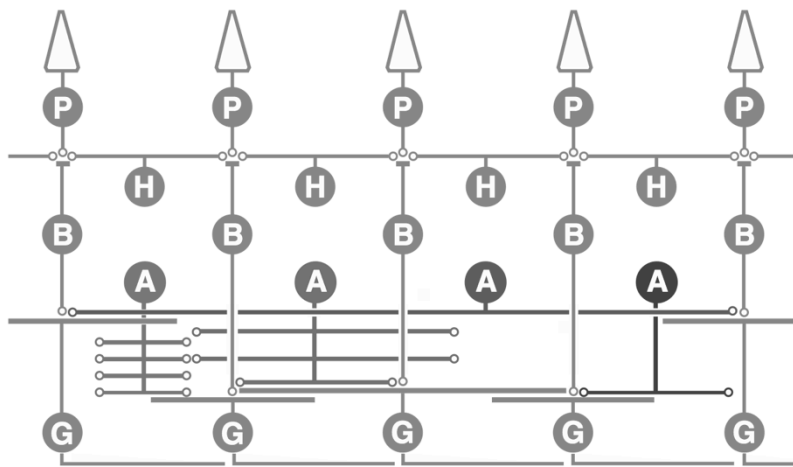
Chichilnisky 2001
Baccus and Meister 2002
Pillow et al 2005, 2008

How well do linear-nonlinear models explain the retina in natural vision?



see also
Heitman et al., 2014

Modeling ganglion cells with convolutional neural networks (CNNs)

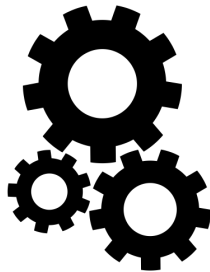


Train the model to minimize the error between predictions and recorded data

Modeling ganglion cells with convolutional neural networks (CNNs)

Challenges

trainability

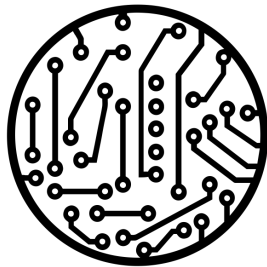


Models are complex, can easily over-fit training data

Modeling ganglion cells with convolutional neural networks (CNNs)

Challenges

neural structure



No reason why the structure or features of learned CNNs would be similar to the retina

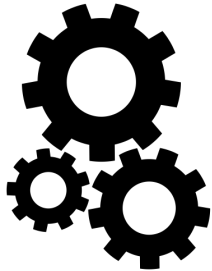
Modeling ganglion cells with convolutional neural networks (CNNs)

Challenges

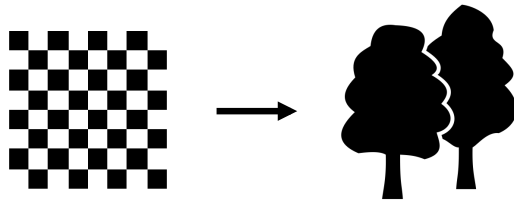
neural function



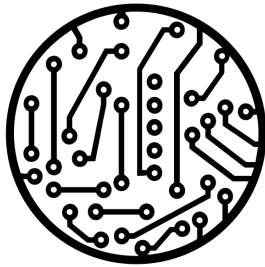
Algorithms identified by the model may not be the same as those used by the retina



CNNs capture substantially more retinal responses than previous models



CNNs generalize better than simpler models

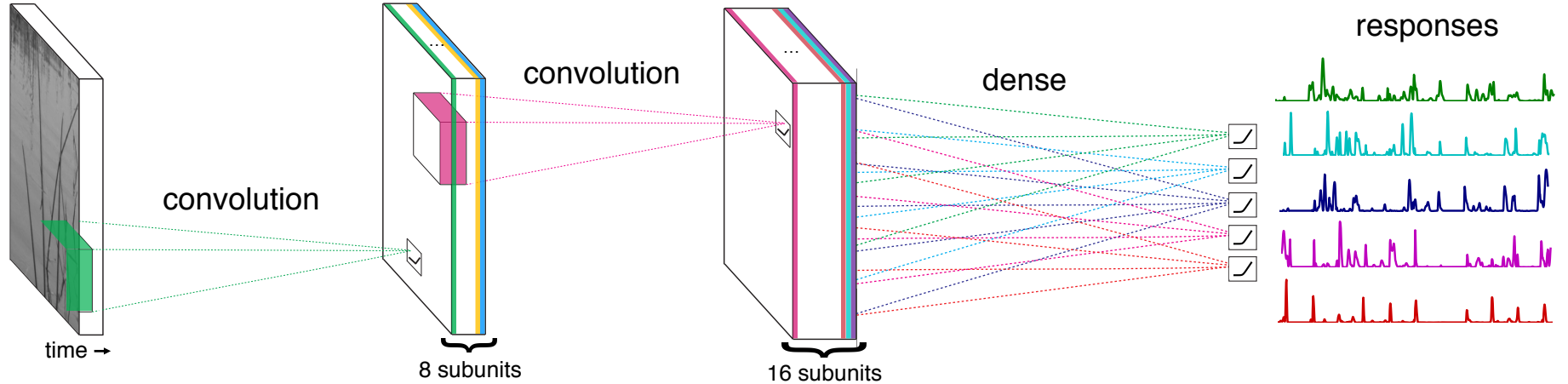


CNN internal units correspond to interneurons in the retinal circuitry



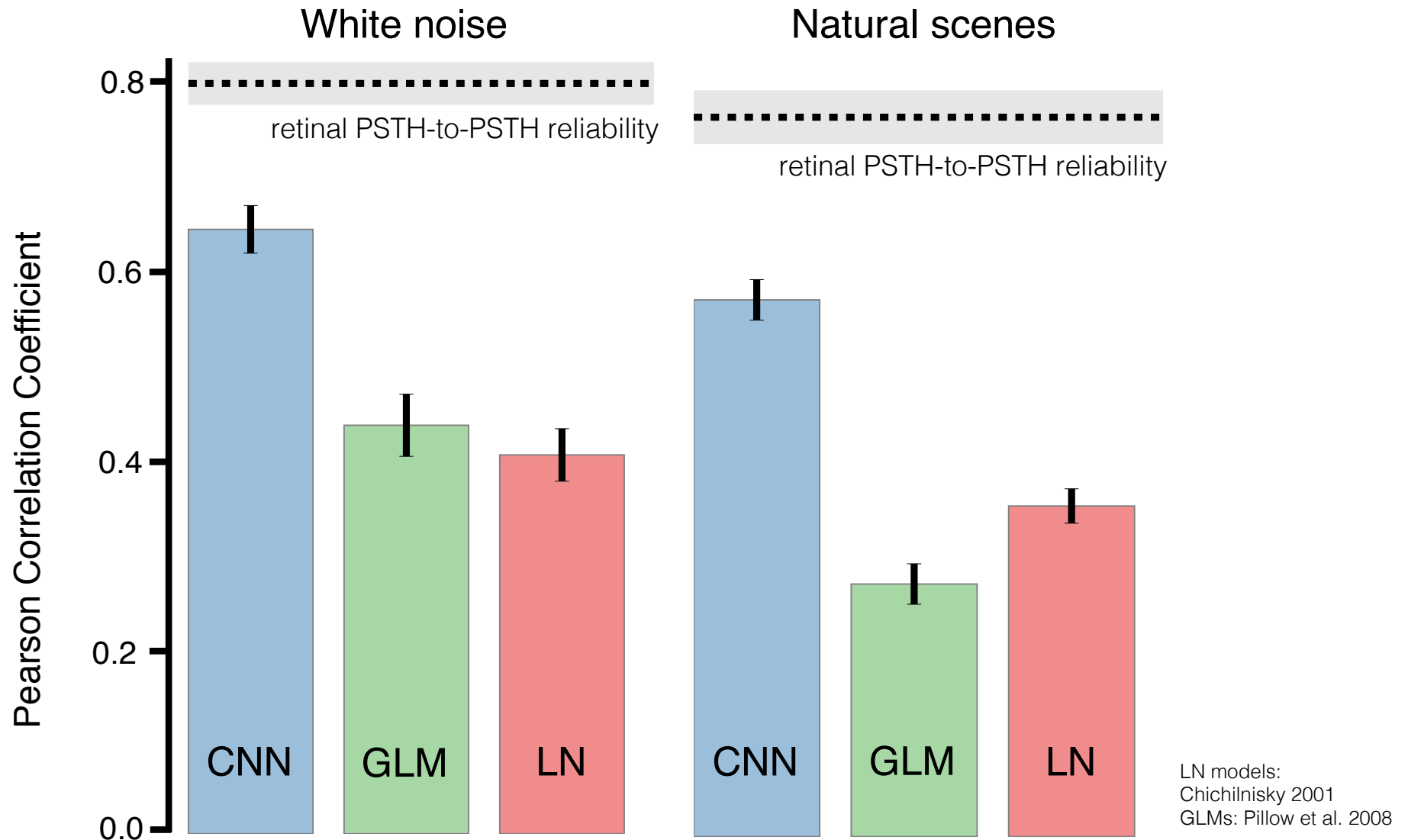
CNNs learn aspects of retinal variability, computation, and adaptation

Convolutional neural network model

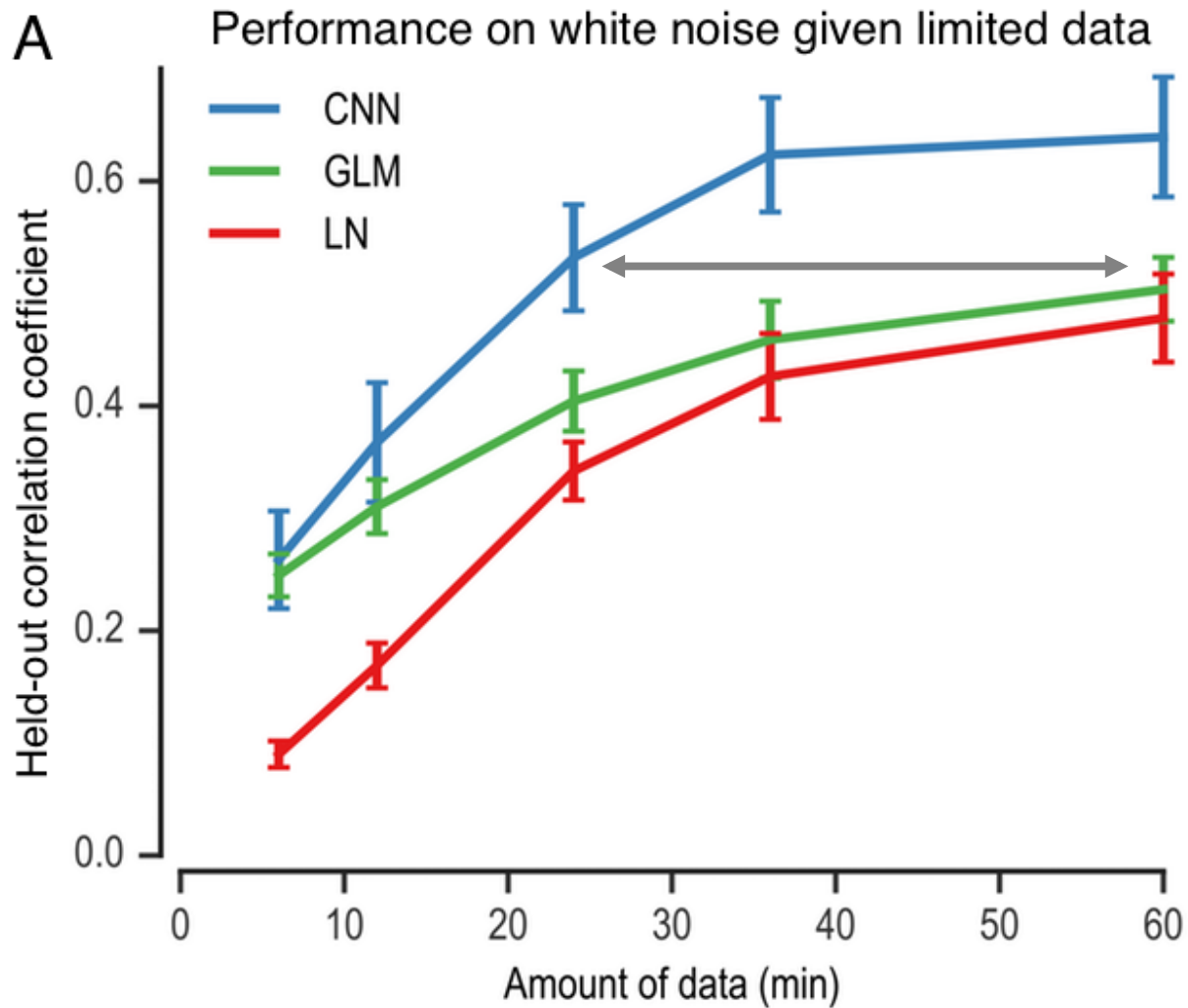


Three layers works best!

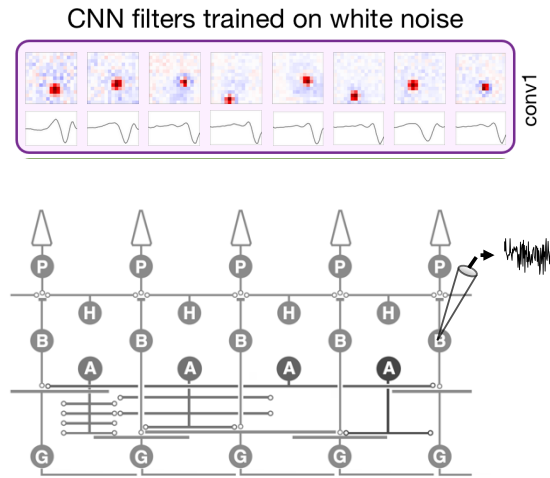
CNNs approach retinal reliability



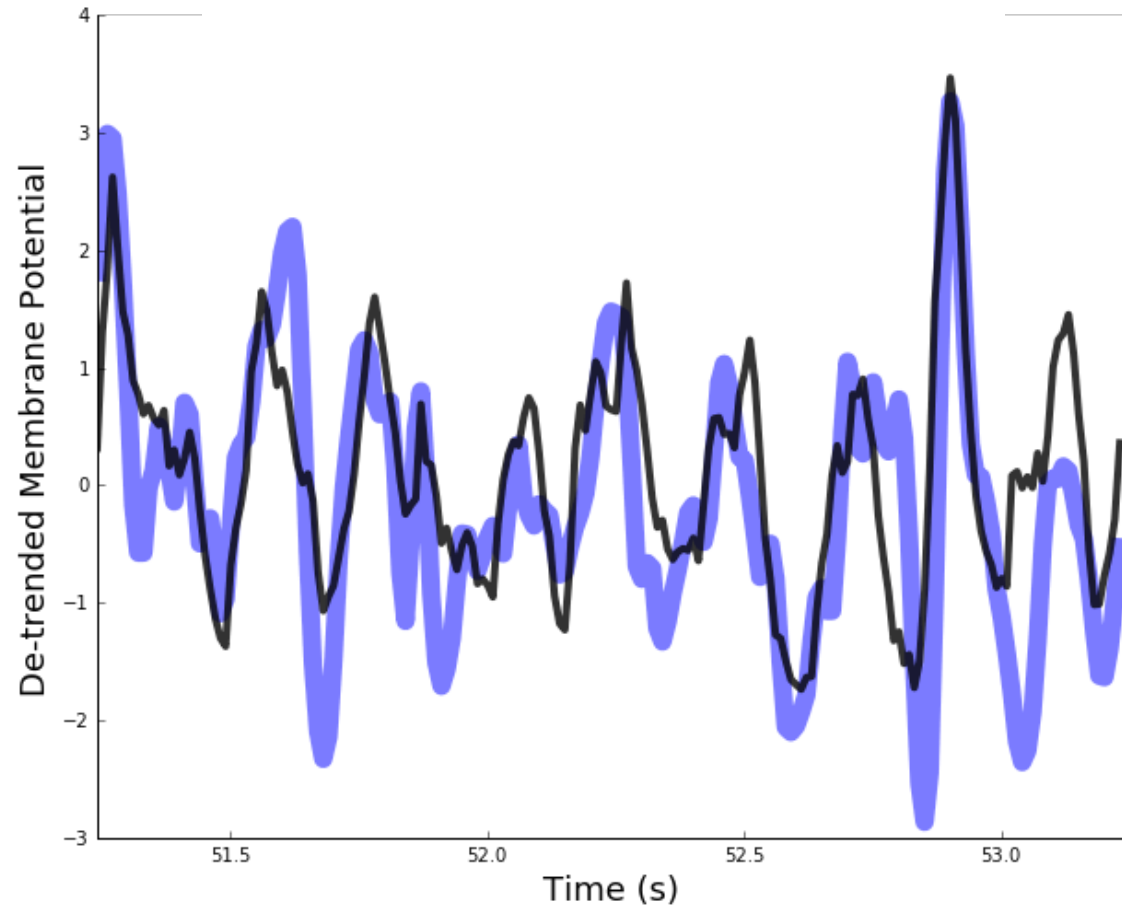
CNN trained on less data outperforms simpler models on more data



Features bear striking resemblance to internal structure in retina

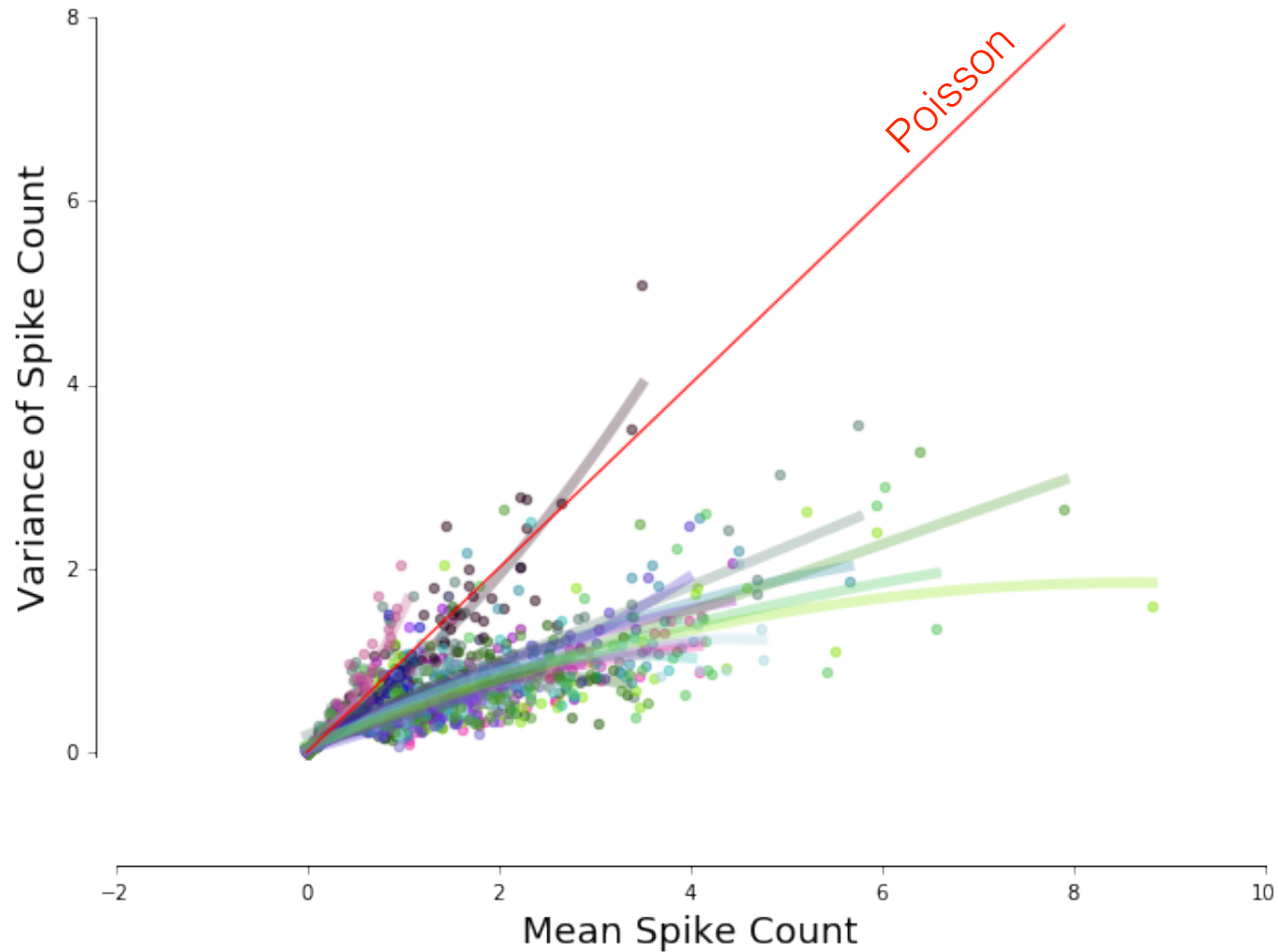


CNN first layer pre-ReLU activity
Bipolar cell membrane potential

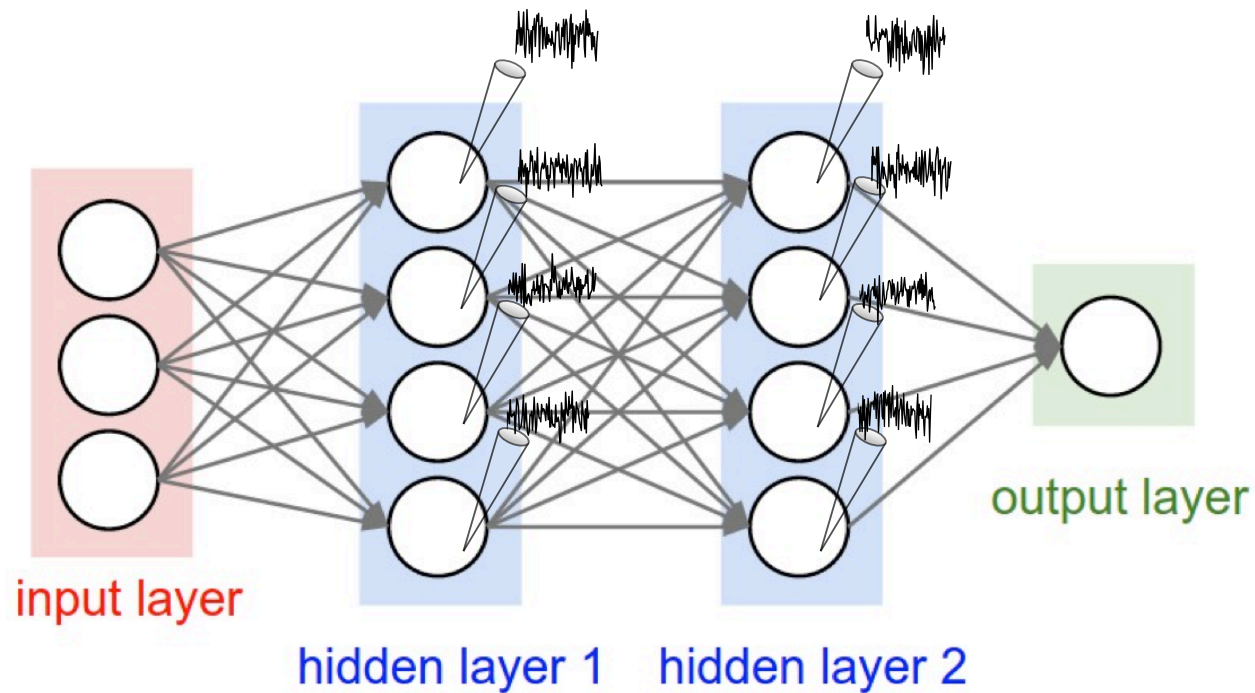


Intracellular data courtesy of
Pablo Jadzinsky and David Kastner

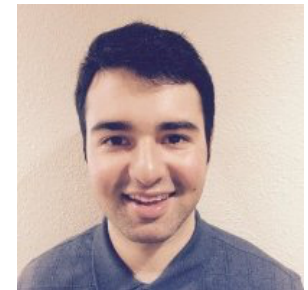
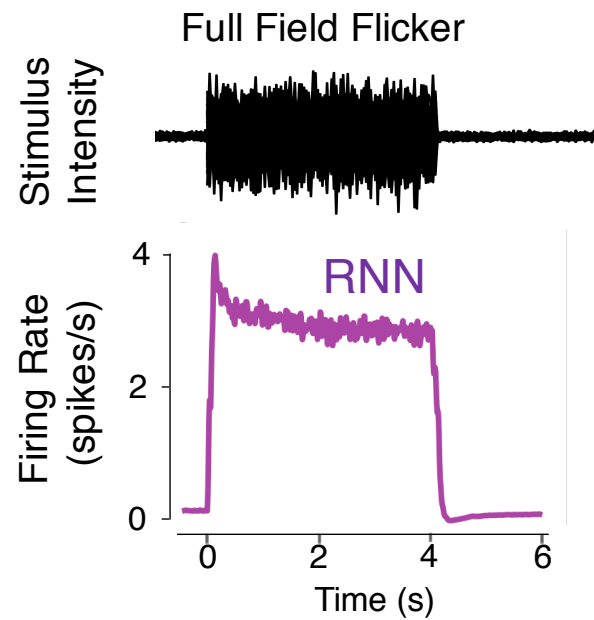
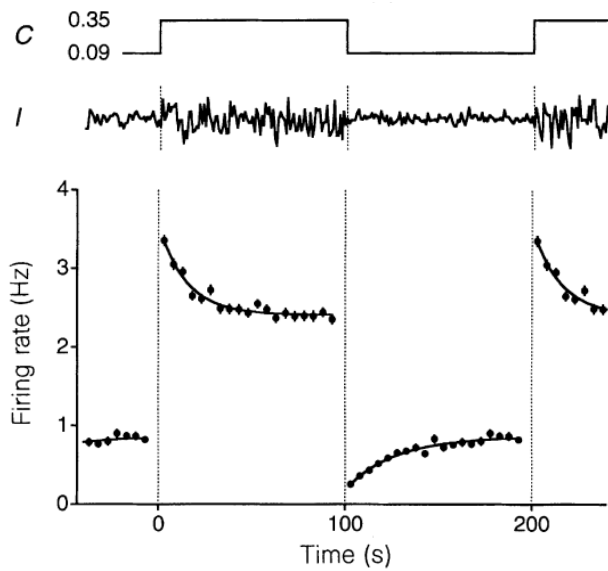
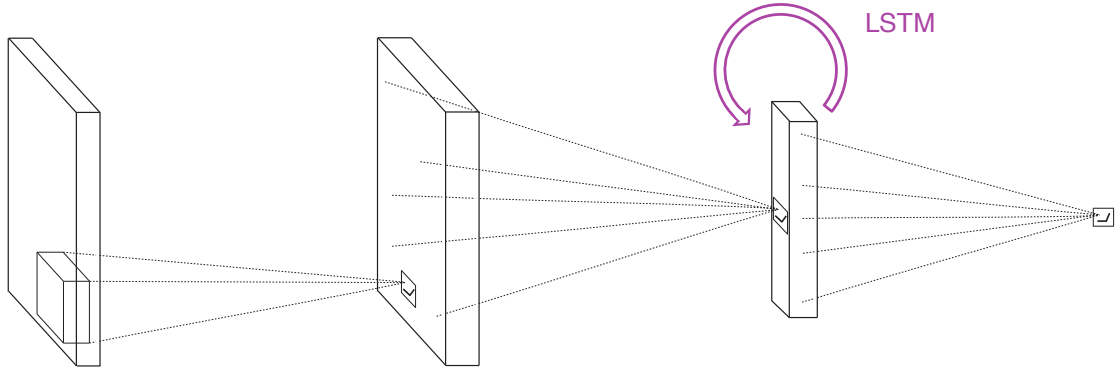
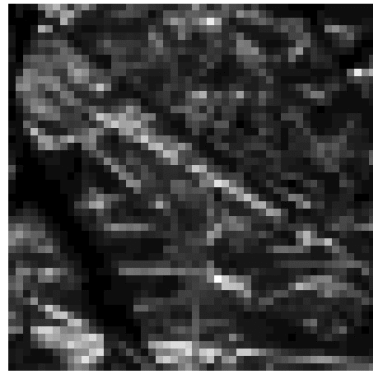
Most retinal neurons have sub-Poisson variability
(while LNP models are Poisson)



We can inject Gaussian noise into each hidden unit of our CNN model

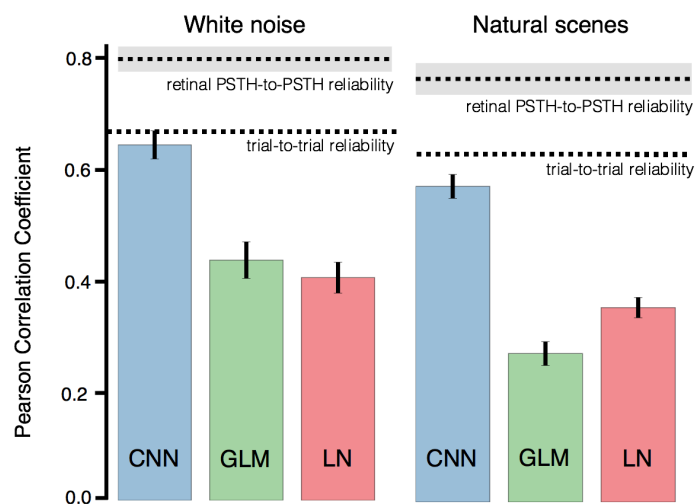


Capturing contrast adaptation from retinal responses to natural scenes



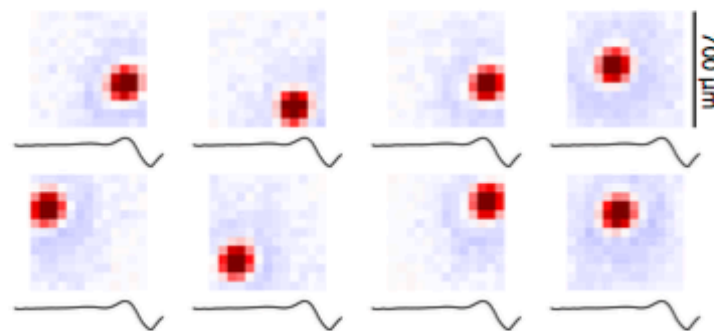
Aran Nayebi

Summary

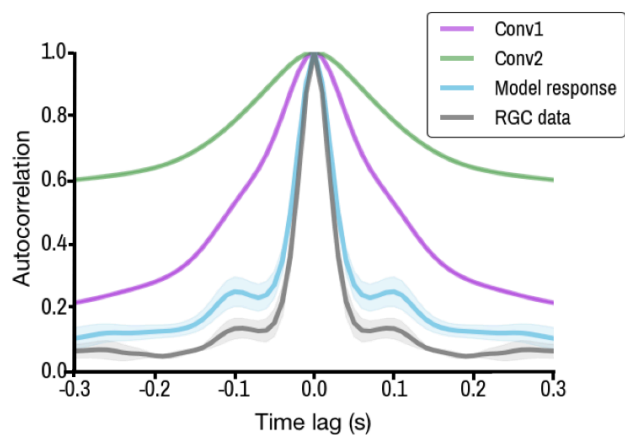


CNNs capture substantially more retinal responses than previous models.

CNNs also generalize better to different stimuli classes.



CNNs learn the internal, nonlinear structure of the retina



We can capture not only the mean response, but also how variability scales with the mean

Our CNN models reproduce principles of signal processing inside retina without having direct access to it!

Talk Outline

- **Applying deep learning to the brain:**
 - Recurrent neural networks for context dependent decision making
 - Feed-forward networks for modeling the ventral visual stream
 - State of the art models of retinal function
- **Theory of deep learning:**
 - Optimization
 - Expressivity
 - Generalization
- **Inspiration from neuroscience back to deep learning:**
 - Canonical cortical microcircuits
 - Nested loop architectures
 - Avoiding catastrophic forgetting through synaptic complexity
 - Learning asymmetric recurrent generative models

Some of the theoretical puzzles of deep learning

Trainability: if a good network solution exists with small training error, how do we find it? And what makes a learning problem difficult?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks ICLR 2014.

A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, CogSci 2013.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

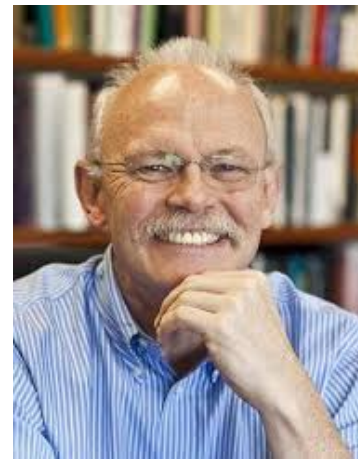
Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

Generalizability: what principles do deep networks use to place probability / make decisions in regions of input space with little data?

M. Advani and S. Ganguli, Statistical Mechanics of Optimal Convex Inference in High Dimensions, Physical Review X, 2016.

Expressiveness, Memorization, Stability, and Flat versus sharp minima.

A Mathematical Theory of Semantic Development*



Joint work with: Andrew Saxe and Jay McClelland

*AKA: The misadventures of an “applied physicist” wandering around the psychology department

What is “semantic cognition”?

Human semantic cognition: Our ability to learn, recognize, comprehend and produce inferences about properties of objects and events in the world, especially properties that are not present in the current perceptual stimulus

For example:

Does a cat have fur?

Do birds fly?

Our ability to do this likely relies on our ability to form internal representations of categories in the world

Psychophysical tasks that probe semantic cognition

Looking time studies: Can an infant distinguish between two categories of objects? At what age?

Property verification tasks: Can a canary move? Can it sing?
Response latency => central and peripheral properties

Category membership queries: Is a sparrow a bird? An ostrich?
Response latency => typical / atypical category members

Inductive generalization:

(A) Generalize familiar properties to novel objects:
i.e. a “blick” has feathers. Does it fly? Sing?

(B) Generalize novel properties to familiar objects:
i.e. a bird has gene “X”. Does a crocodile have gene X?
Does a dog?

The project that really keeps me up at night



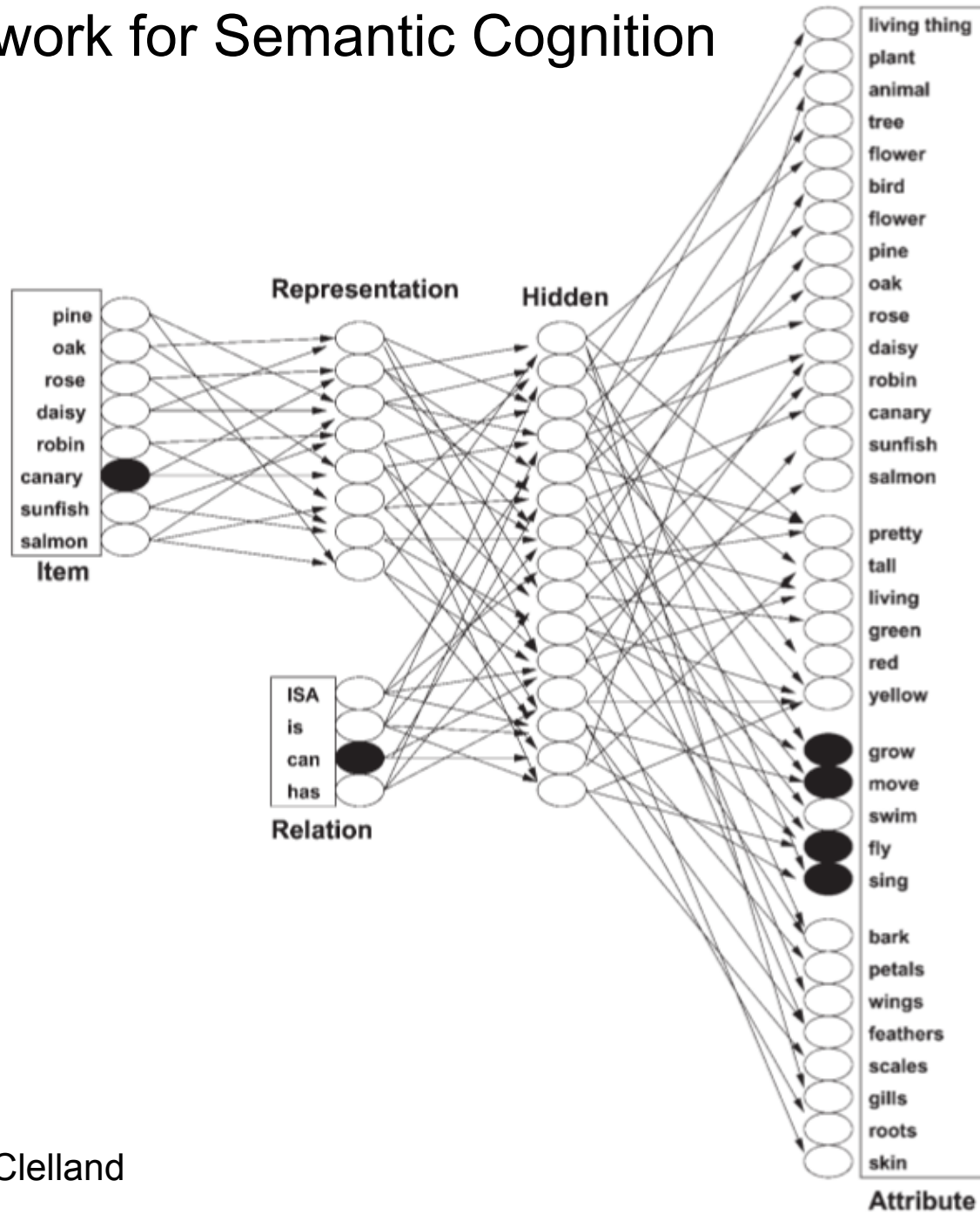
Semantic Cognition Phenomena

Rogers & McClelland: Précis of *Semantic Cognition*

Table 1. Six key phenomena in the study of semantic abilities

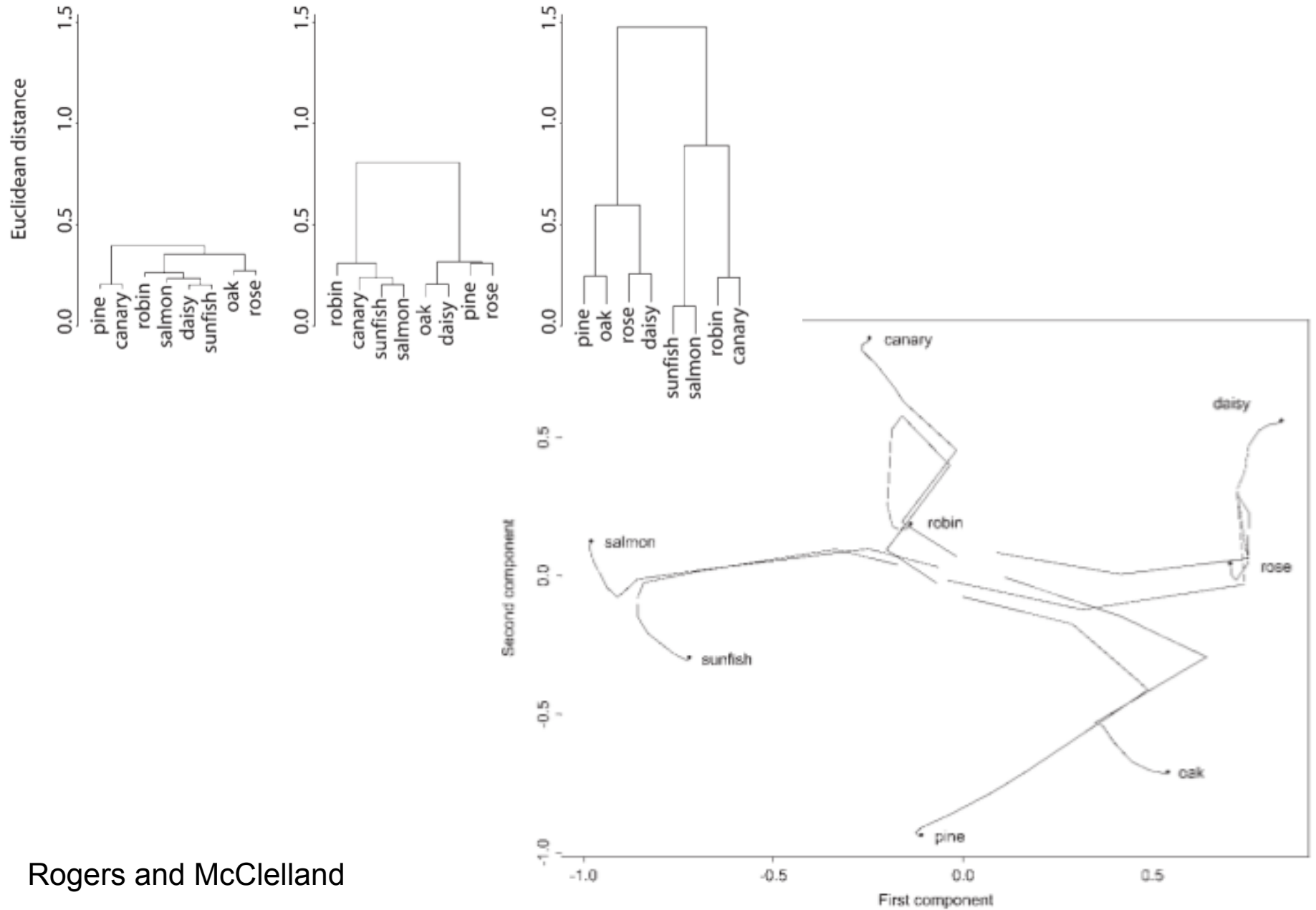
Phenomenon	Example
Progressive differentiation of concepts	Children acquire broader semantic distinctions earlier than more fine-grained distinctions. For example, when perceptual similarity among items is controlled, infants differentiate animals from furniture around 7–9 months of age, but do not make finer-grained distinctions (e.g., between fish and birds or chairs and tables) until somewhat later (Pauen 2002a; Mandler et al. 1991); and a similar pattern of coarse-to-fine conceptual differentiation can be observed between the ages of 4 and 10 in verbal assessments of knowledge about which predicates can appropriately apply to which nouns (Keil 1989).
Category coherence	Some groupings of objects (e.g., “the set of all things that are dogs”) seem to provide a useful basis for naming and inductive generalization, whereas other groupings (e.g., “the set of all things that are blue”) do not. How does the semantic system “know” which groupings of objects should be used for purposes of naming and inductive generalization, and which should not?
Domain-specific attribute weighting	Some properties seem of central importance to a given concept, whereas others do not. For instance, “being cold inside” seems important to the concept <i>refrigerator</i> , whereas “being white” does not. Furthermore, properties that are central to some concepts may be unimportant for others – although having a white color may seem unimportant for <i>refrigerator</i> , it seems more critical to the concept <i>polar bear</i> . What are the mechanisms that support domain-specific attribute weighting?
Illusory correlations	Children and adults sometimes attest to beliefs that directly contradict their own experience. For example, when shown a photograph of a kiwi bird – a furry-looking animal with eyes but no discernible feet – children may assert that the animal can move “because it has feet,” even while explicitly stating that they can see no feet in the photograph. Such illusory correlations appear to indicate some organizing force behind children’s inferences that goes beyond “mere” associative learning. What mechanisms promote illusory correlations?
Conceptual reorganization	Children’s inductive projection of biological facts to various different plants and animals changes dramatically between the ages of 4 and 10. For some researchers, these changing patterns of induction indicate changes to the implicit theories that children bring to bear on explaining biological facts. What mechanism gives rise to changing induction profiles over development?
The importance of causal knowledge	A variety of evidence now indicates that, in various kinds of semantic induction tasks, children and adults strongly weight causally central properties over other salient but non-causal properties. Why are people sensitive to causal properties?

A Network for Semantic Cognition



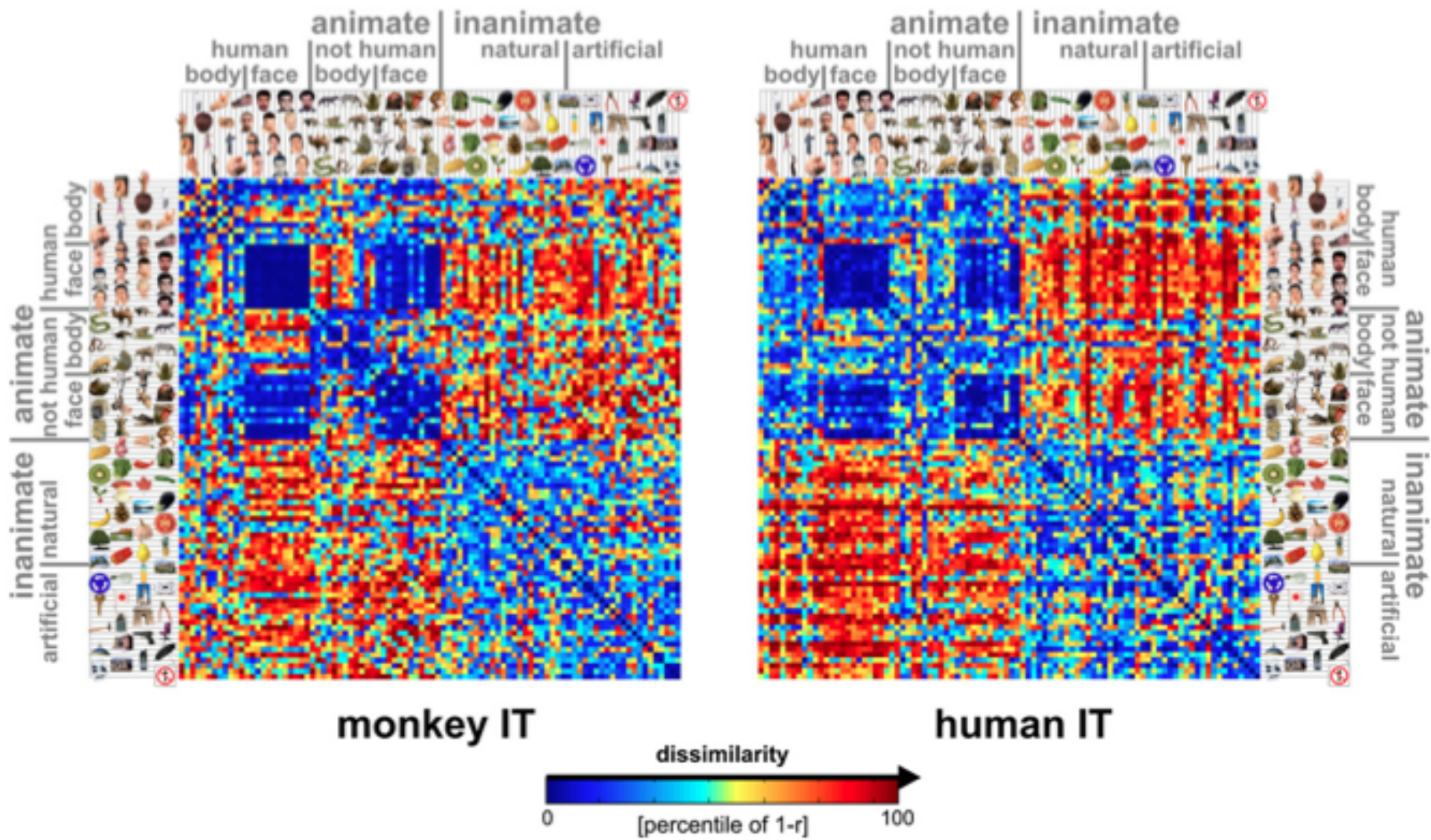
Rogers and McClelland

Evolution of internal representations

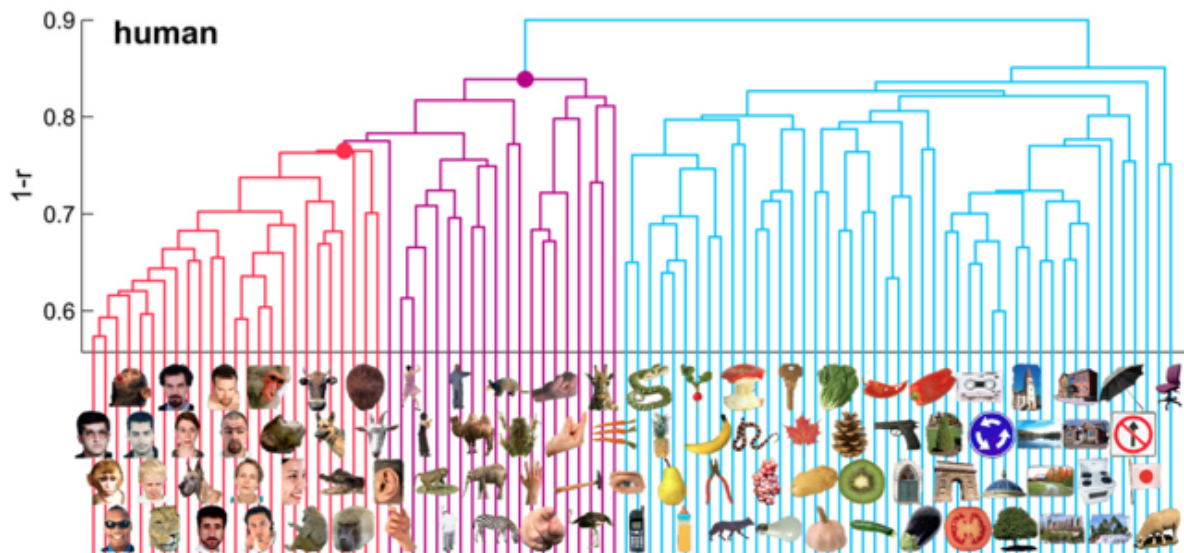
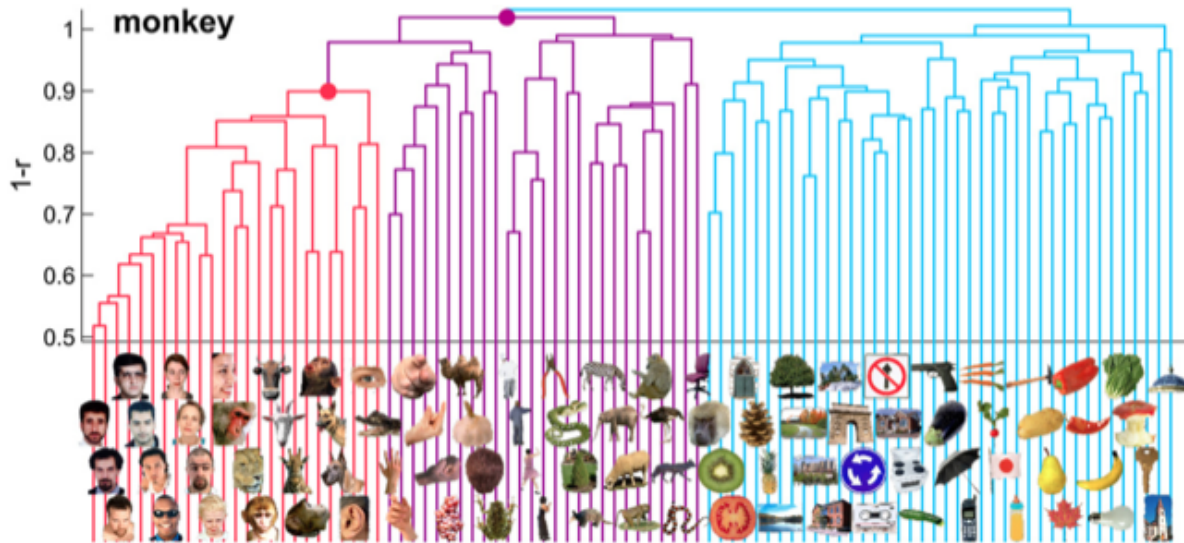


Rogers and McClelland

Categorical representations in human and monkey



Categorical representations in human and monkey



Evolution of internal representations

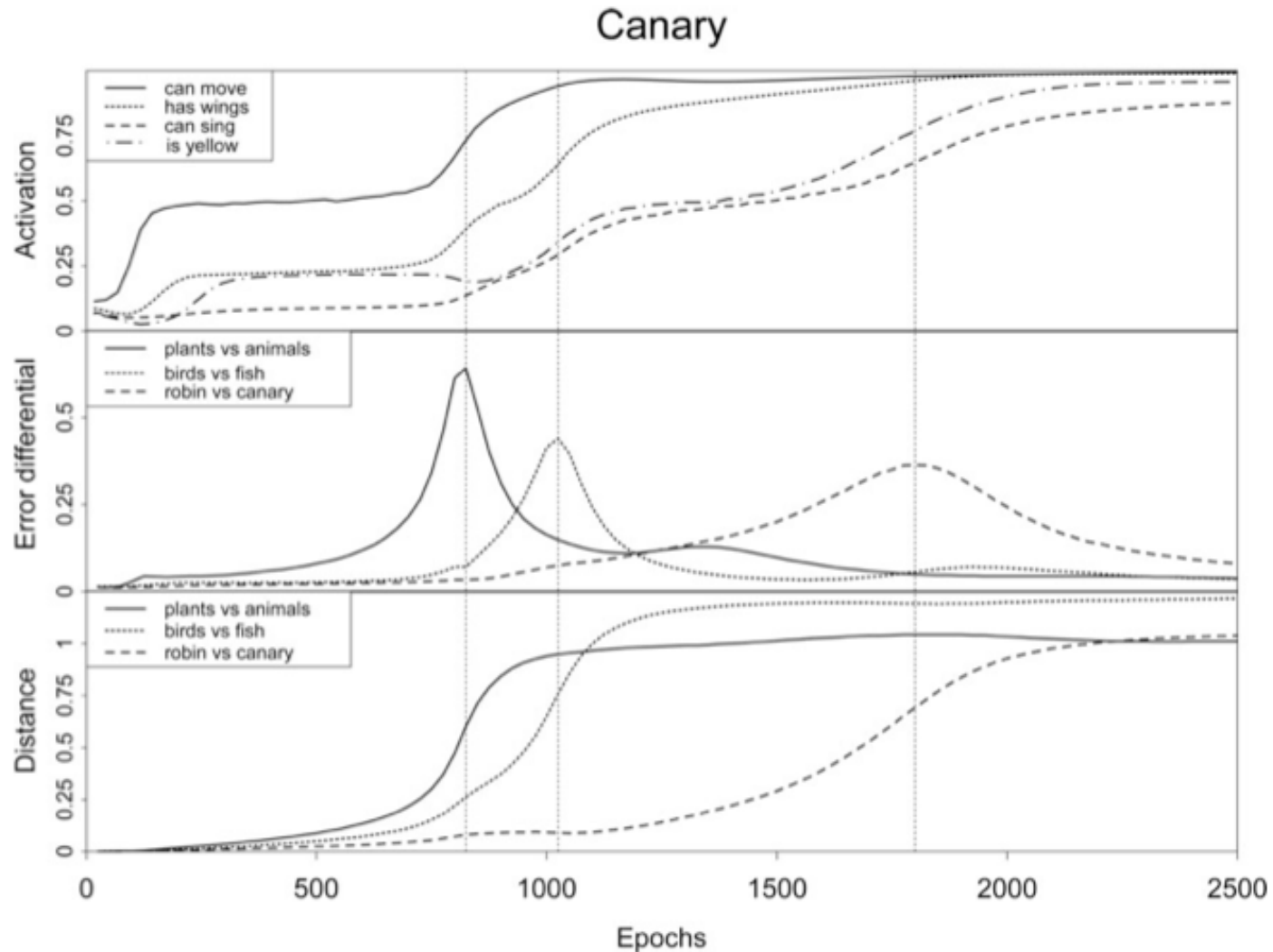


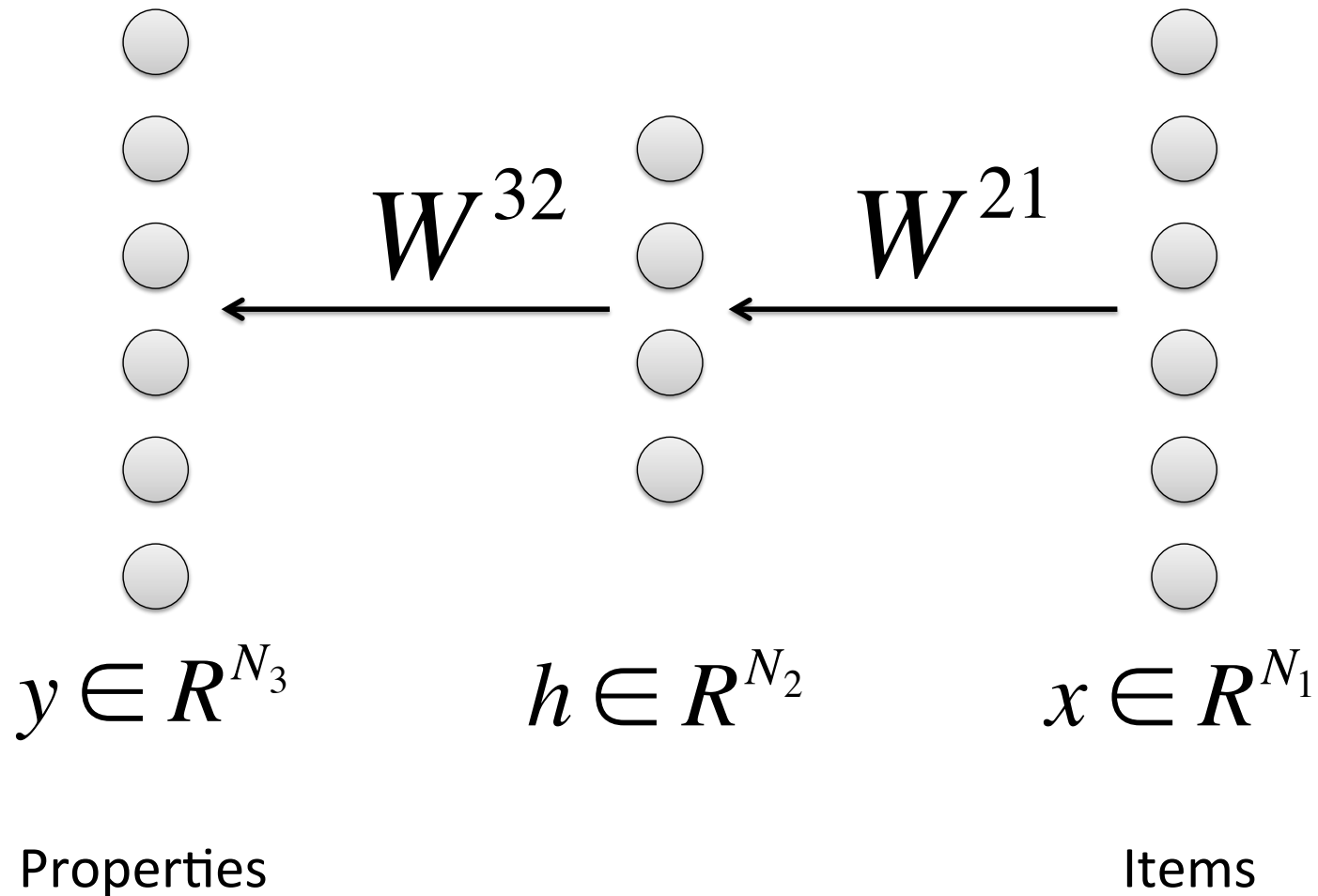
Figure 5. **Bottom:** Mean Euclidean distance between plants and animals, birds and fish, and canary and robin internal representations throughout training. **Middle:** Average magnitude of the error signal propagating back from properties that reliably discriminate plants from animals, birds from fish, or the canary from the robin, at different points throughout training when the model is presented with the canary as input. **Top:** Activation of a property shared by animals (*can move*) or birds (*can fly*), or unique to the canary (*can sing*), when the model is presented with the input canary can at different points throughout training.

Theoretical questions

- What are the mathematical principles underlying the hierarchical self-organization of internal representations in the network?
- What are the relative roles of:
 - nonlinear input-output response
 - learning rule
 - input statistics (second order? higher order?)
- What is a mathematical definition of category coherence, and How does it relate the speed of category learning?
- Why are some properties learned more quickly than others?
- How can we explain changing patterns of inductive generalization over developmental time scales?

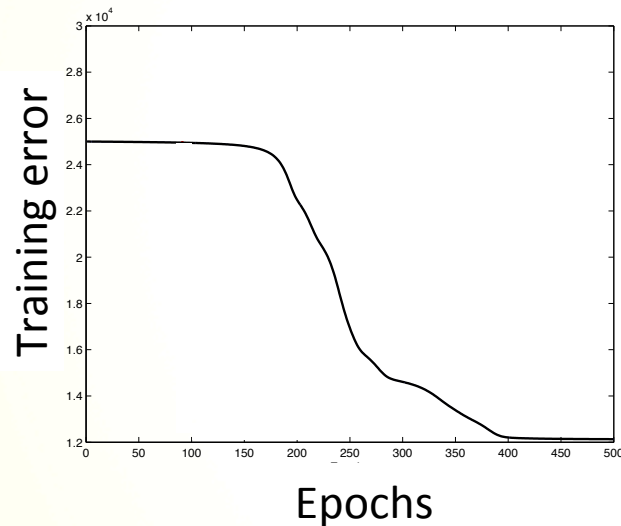
Problem formulation

We analyze a fully linear three layer network $y = W^{32}W^{21}x$

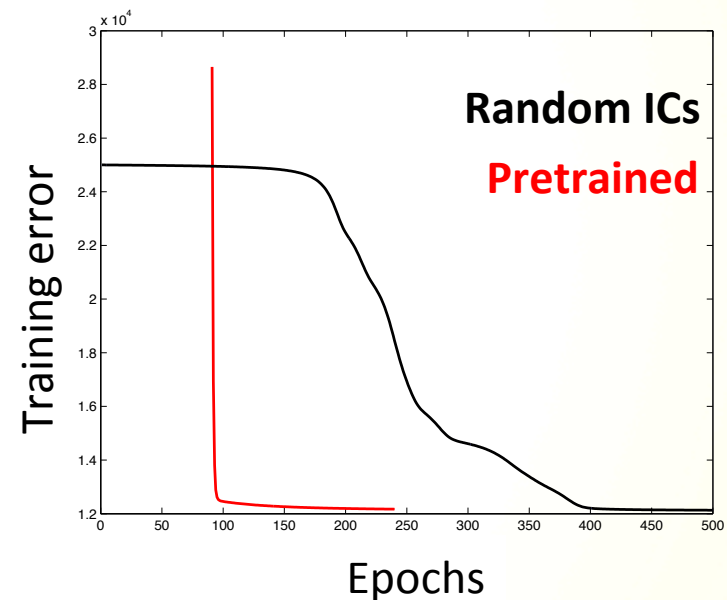


Nontrivial learning dynamics

Plateaus and sudden transitions



Faster convergence from pretrained initial conditions



- Build intuitions for nonlinear case by analyzing linear case

Learning dynamics

- Network is trained on a set of items and their properties

$$\{x^\mu, y^\mu\}, \mu = 1, \dots, P.$$

- Weights adjusted using standard backpropagation:
 - Change each weight to reduce the error between desired network output and current network output

$$\Delta W^{21} = \lambda W^{32T} (y^\mu - \hat{y}^\mu) x^{\mu T}$$

$$\Delta W^{32} = \lambda (y^\mu - \hat{y}^\mu) h^{\mu T}$$

- Highlights the error-corrective aspect of this learning process

Learning dynamics

In linear networks, there is an equivalent formulation that highlights the role of the statistics of the training environment:

$$\begin{aligned} \text{Input correlations:} & \quad \Sigma^{11} \equiv E[xx^T] \\ \text{Input-output correlations:} & \quad \Sigma^{31} \equiv E[yx^T] \end{aligned}$$

Equivalent dynamics:

$$\begin{aligned} \tau \frac{d}{dt} W^{21} &= W^{32T} (\Sigma^{31} - W^{32} W^{21} \Sigma^{11}) \\ \tau \frac{d}{dt} W^{32} &= (\Sigma^{31} - W^{32} W^{21} \Sigma^{11}) W^{21T} \end{aligned}$$

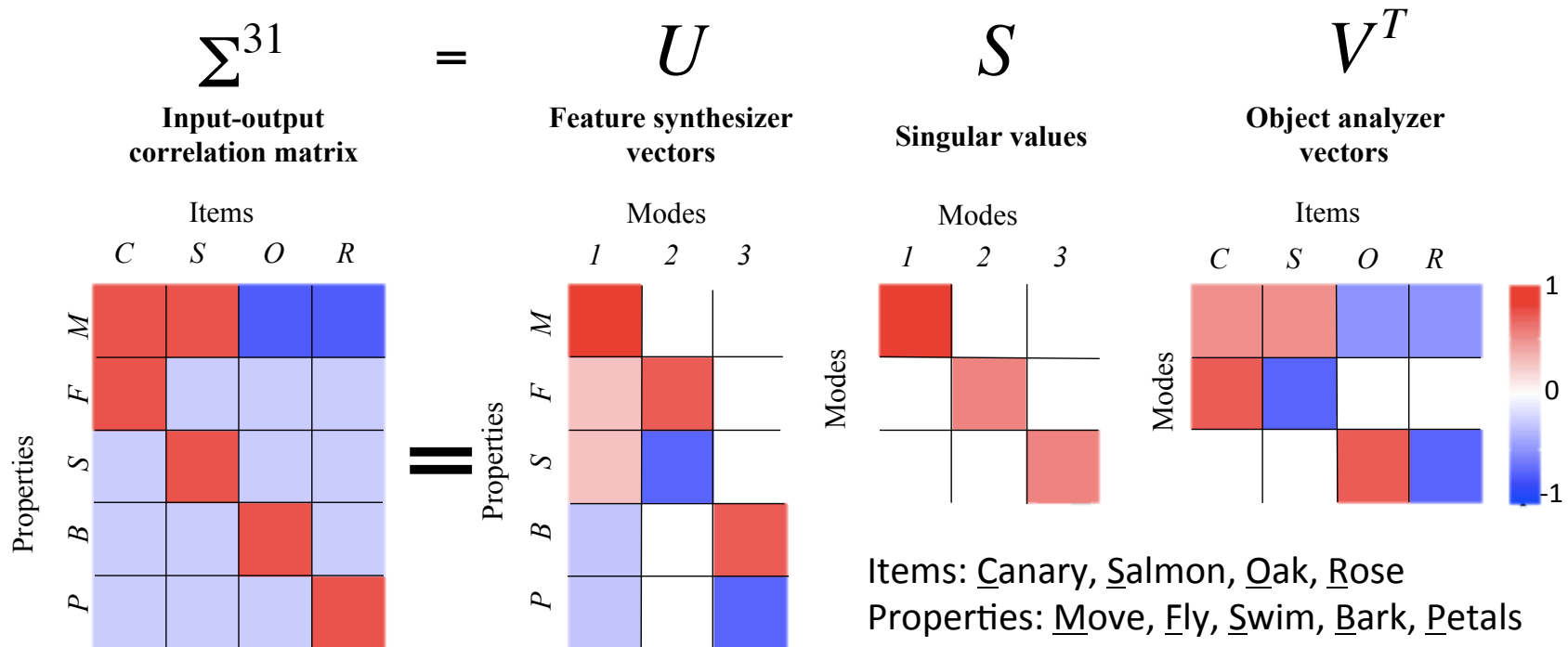
- Learning driven only by correlations in the training data
- Equations coupled and nonlinear

Decomposing input-output correlations

The learning dynamics can be expressed using the SVD of Σ^{31}

$$\Sigma^{31} = U^{33} S^{31} V^{11T} = \sum_{\alpha=1}^{N_1} s_{\alpha} u^{\alpha} v^{\alpha T}$$

Mode α links a set of coherently covarying properties u^{α} to a set of coherently covarying items $v^{\alpha T}$ with strength s_{α}



Analytical learning trajectory

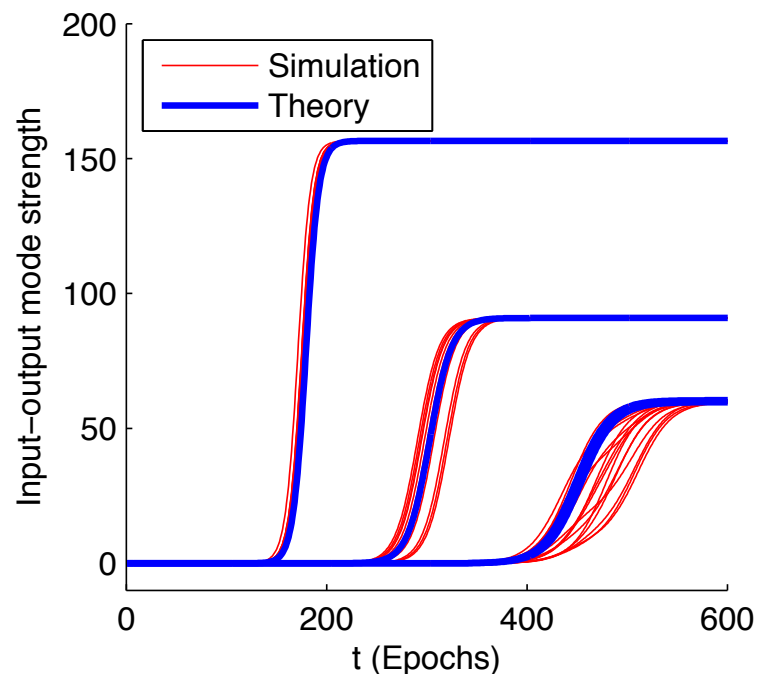
The network's input-output map is exactly

$$W^{32}(t)W^{21}(t) = \sum_{\alpha=1}^{N_2} \frac{a(t, s_{\alpha}, a_{\alpha}^0) u^{\alpha} v^{\alpha T}}{s e^{2st/\tau}}$$

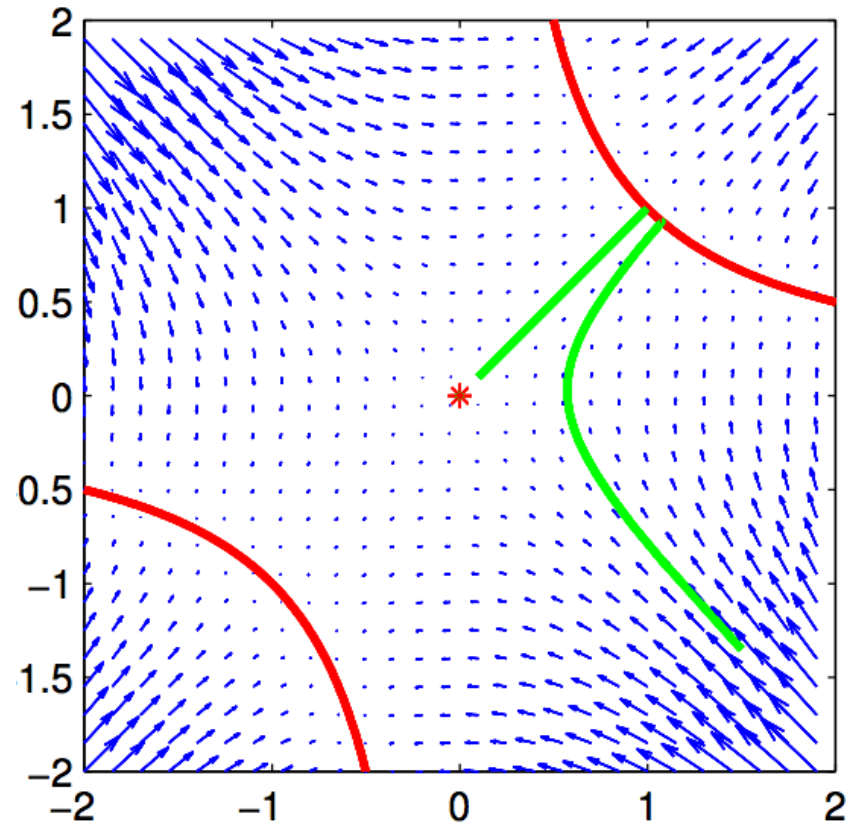
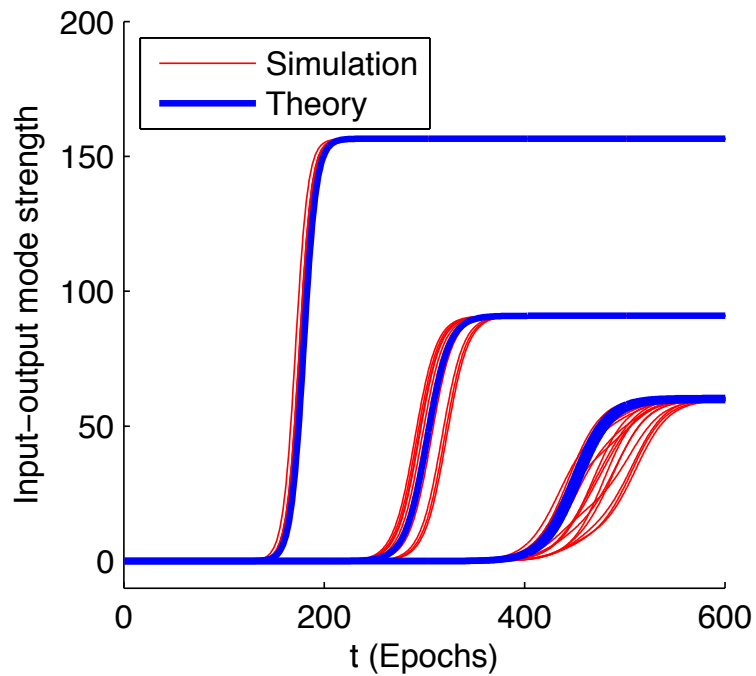
where $a(t, s, a_0) = \frac{e^{2st/\tau} - 1 + s/a_0}{e^{2st/\tau}}$

for a special class of initial conditions and $\Sigma^{11} = I$.

- Each mode evolves independently
- Each mode is **learned in time** $O(\tau/s)$



Origin of the rapid learning transition: saddle point dynamics in synaptic weight space



Take home messages, so far:

Stronger statistical structure

is learned faster!

Strength of structure:

Learning Time

Singular value

$1 / \text{Singular value}$

(Singular vectors:
object analyzers and
feature synthesizers)

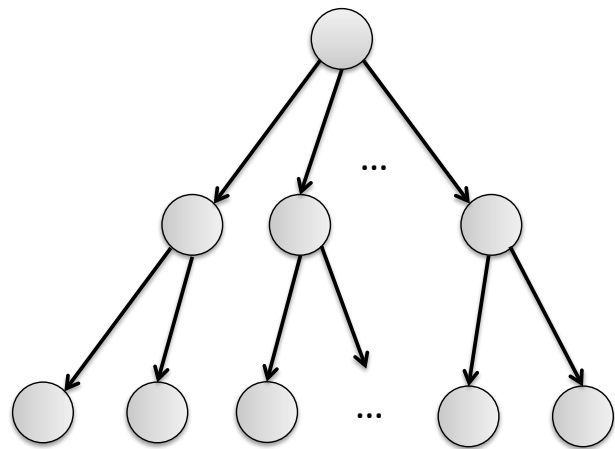
Next: what does all this have to do with the hierarchical
Differentiation of concepts?

Learning hierarchical structure

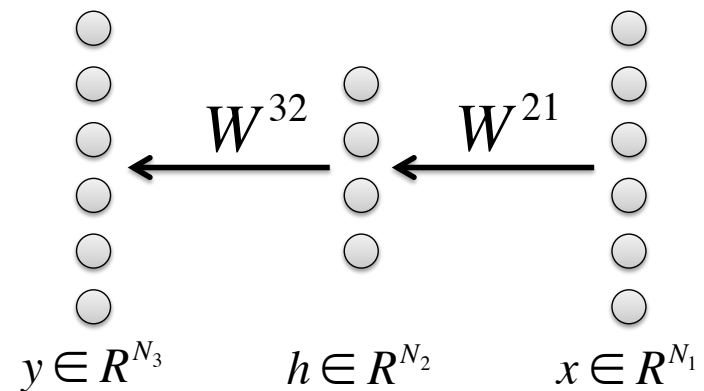
- The preceding analysis describes dynamics in response to a **specific** dataset
- Can we move beyond specific datasets to **general** principles when a neural network is exposed to hierarchical structure?
- We consider training a neural network with data generated by a **hierarchical generative model**

Connecting hierarchical generative models and neural network learning

World



Agent



$\{x^\mu, y^\mu\}, \mu = 1, \dots, P.$

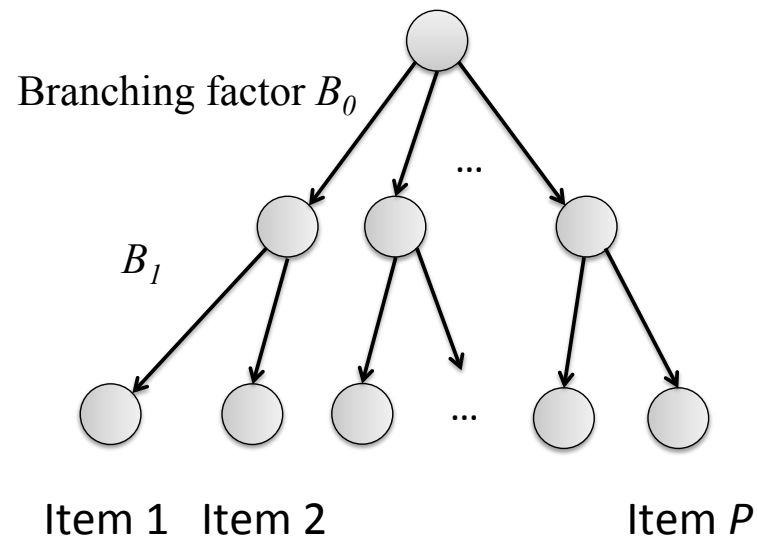


A hierarchical branching diffusion process

Generative model defined
by a tree of nested
categories

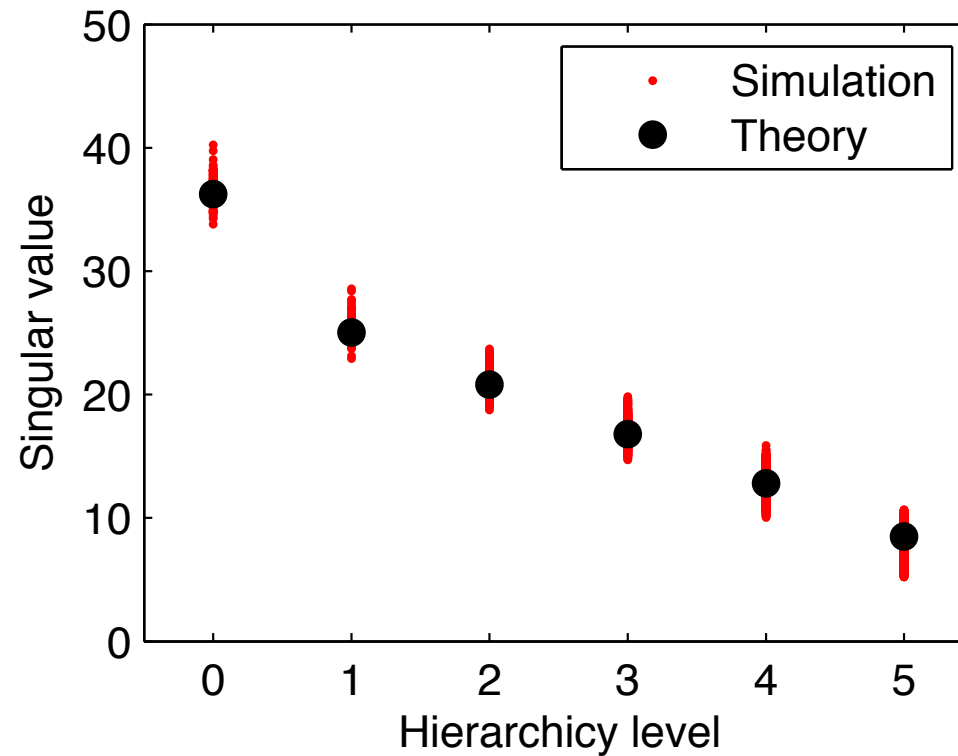
Feature values diffuse
down tree with small
probability ε of changing
along each link

Sampled independently
 N times to produce
 N features



Singular values

The singular values are a ***decreasing*** function of the hierarchy level.



Progressive differentiation

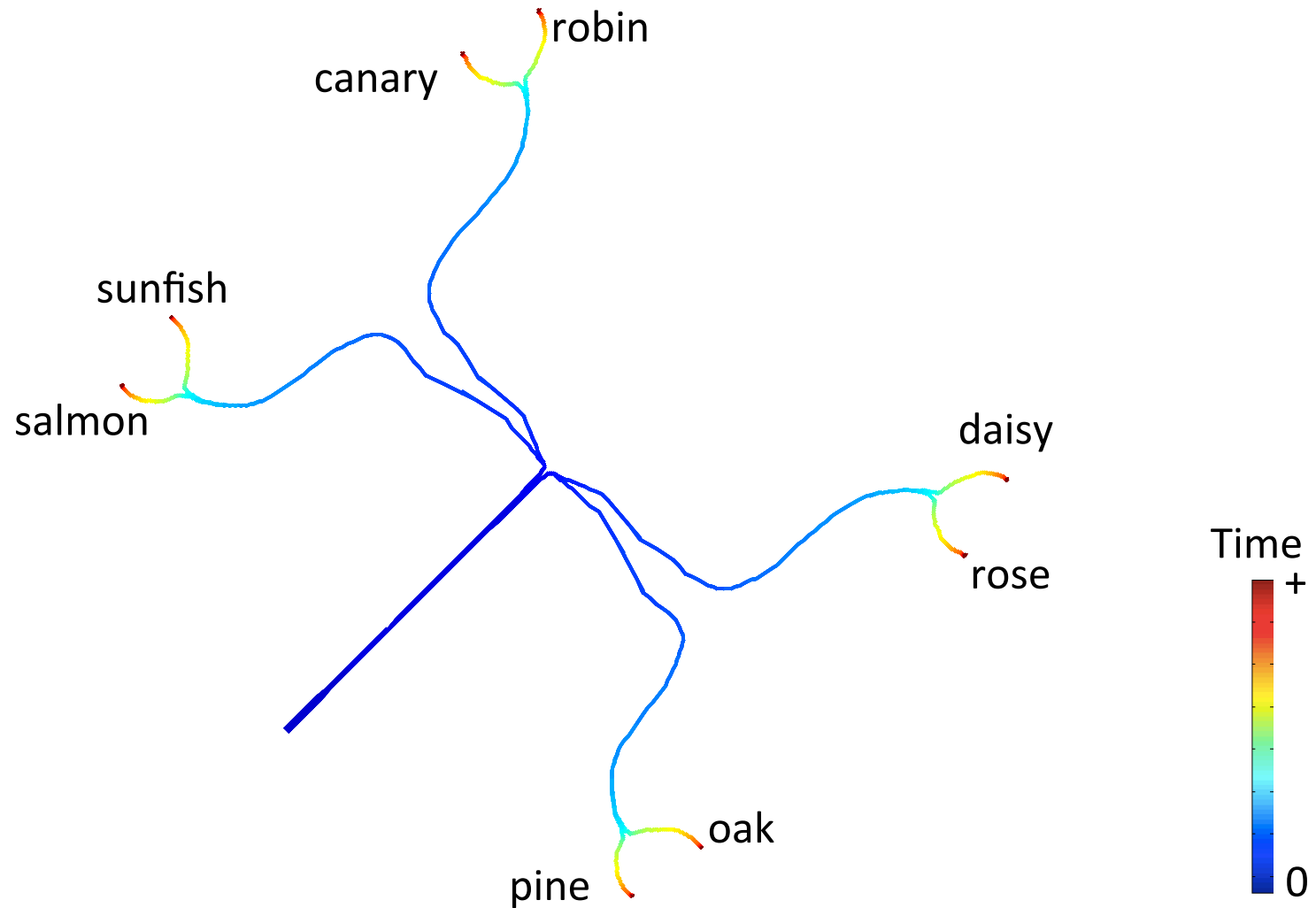
Hence the network **must** exhibit progressive differentiation on **any** dataset generated by this class of hierarchical diffusion processes:

- Network learns input-output modes in time

$$O(\tau/s)$$

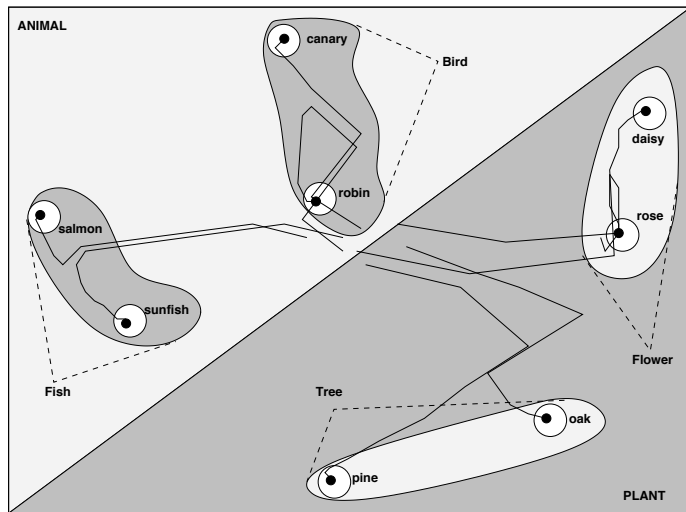
- Singular values of broader hierarchical distinctions are larger than those of finer distinctions
- Input-output modes correspond exactly to the hierarchical distinctions in the underlying tree

Progressive differentiation



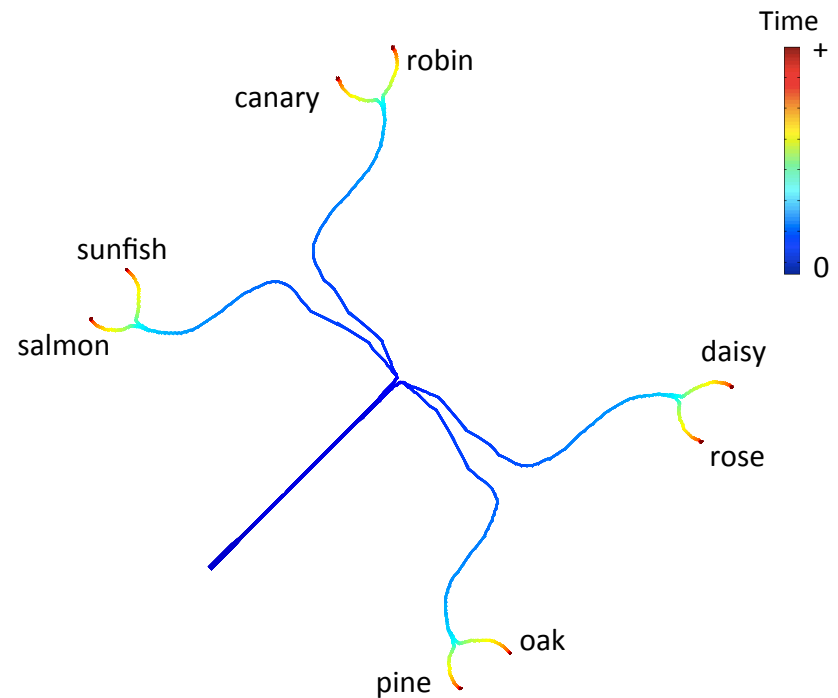
Progressive differentiation

Simulation



Rogers & McClelland, 2004

Analytics



Conclusion

- **Progressive differentiation of hierarchical structure** is a general feature of learning in deep neural networks
- Deep (but not shallow) networks exhibit **stage-like transitions** during learning
- Second order statistics of data are sufficient to drive hierarchical differentiation

Other work

Can analytically understand design principles governing many phenomena previously simulated

- Illusory correlations early in learning
- Familiarity and typicality effects
- Inductive property judgments
- 'Distinctive' feature effects
- Basic level effects
- Category coherence
- Perceptual correlations
- Practice effects

Our framework **connects probabilistic models** and **neural networks**, analytically linking structured environments to learning dynamics.

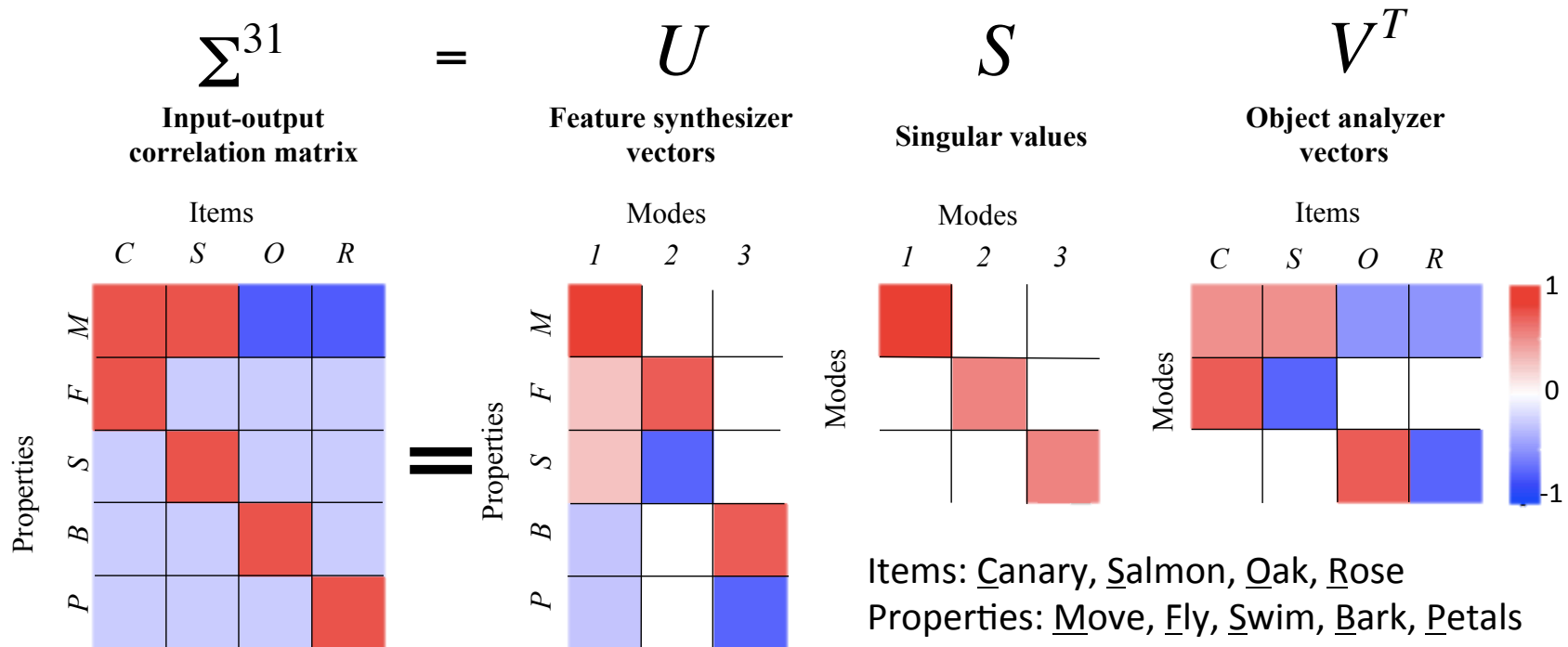
Why are some properties distinctive, or learned faster?

A property = vector across items

An object analyzer = vector across items

If a property is similar to an object analyzer with large singular value then (and only then) will it be learned quickly.

That property is distinctive for the category associated with that object analyzer (i.e. move for animals versus plant)

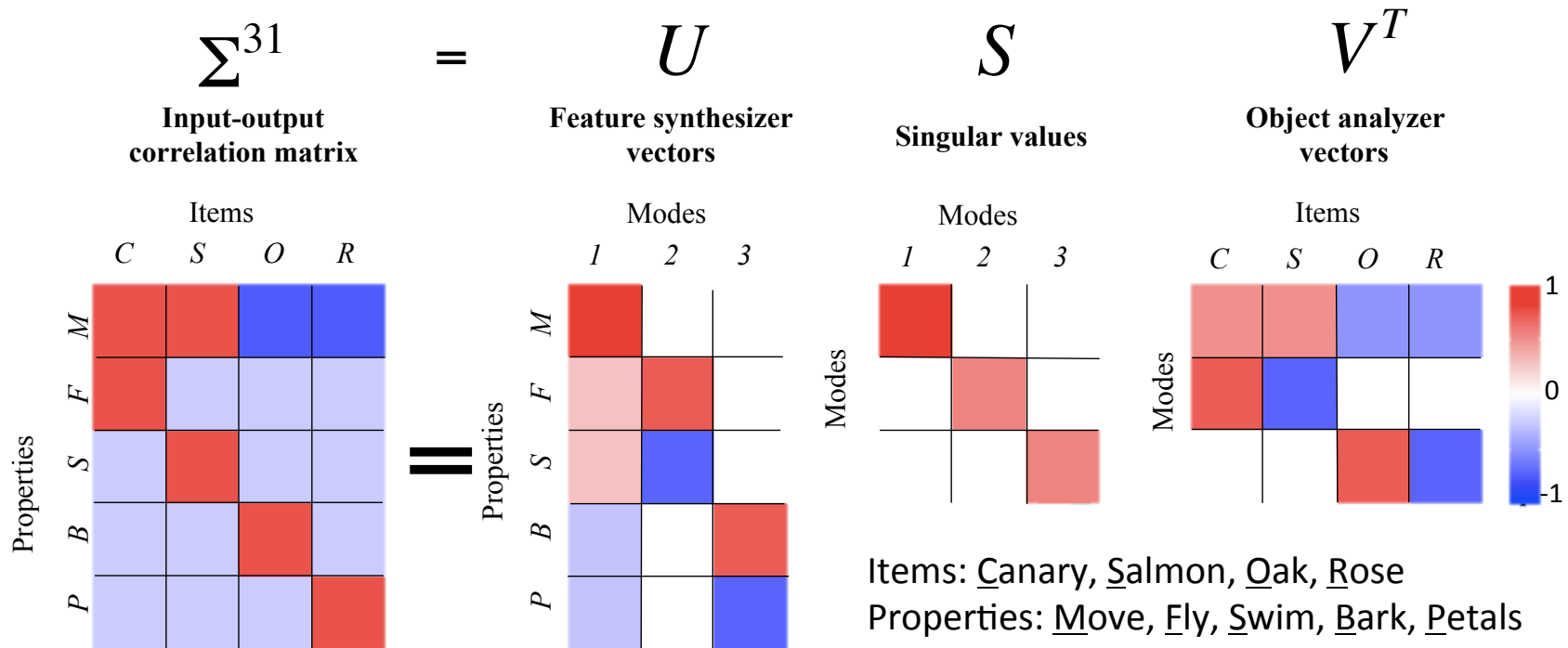


Why are some items more typical members of a category? (i.e. sparrow versus ostrich for the category bird)

An item = vector across properties
 A category feature synthesizer = vector across properties

If an item is similar to the feature synthesizer for a category, then it is a typical member of that category.

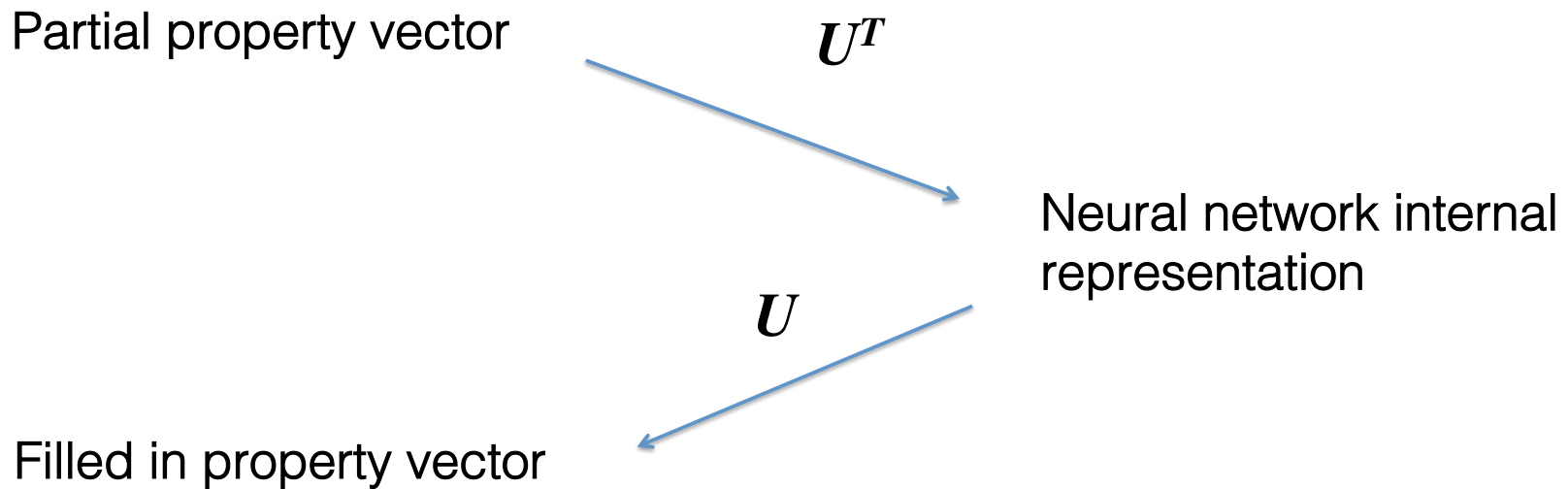
Category membership verification easier for typical versus atypical items.



How is inductive generalization achieved by neural networks? Inferring familiar properties of a novel item.

Given a new partially described object = vector across subset of properties
What are the rest of the object's properties?

i.e. a "blick" has feathers. Does it fly? Sing?



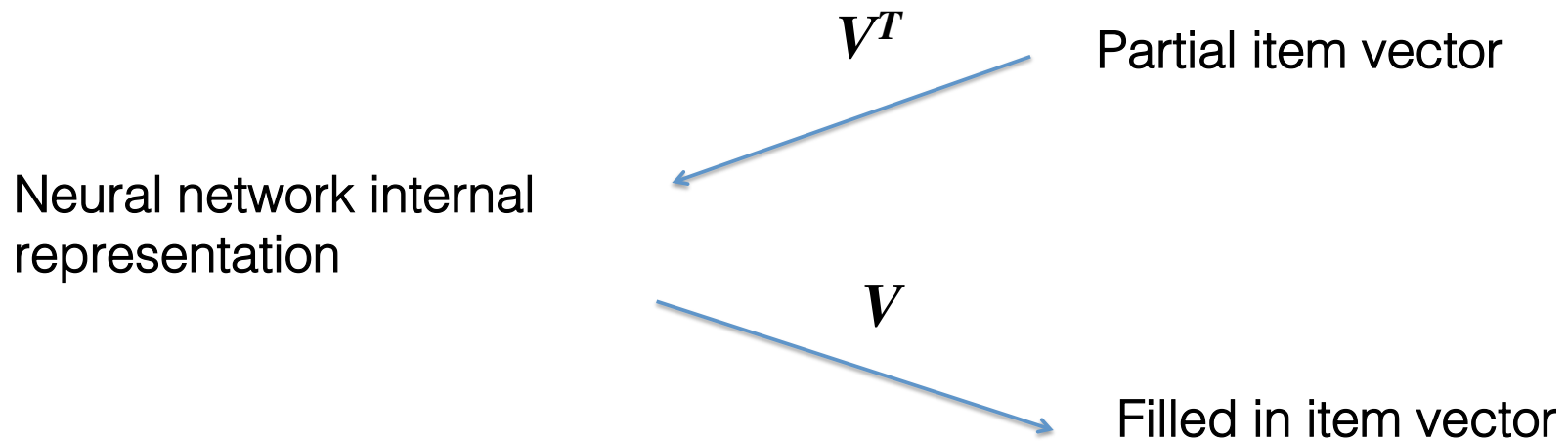
$$\Sigma^{31} = U S V^T$$

Input-output correlation matrix Feature synthesizer vectors Singular values Object analyzer vectors

How is inductive generalization achieved by neural networks? Inferring which familiar objects have a novel property.

Given a new property = vector across subset of items
 Which other items have this property?

i.e. A bird has gene X. Does a crocodile? A dog?



$$\Sigma^{31} = U S V^T$$

Input-output correlation matrix = Feature synthesizer vectors Singular values Object analyzer vectors

What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?

Identify the objects that belong to the category

But must know which features are important for the category

Identify the features that are important for the category

But must know which objects belong to the category

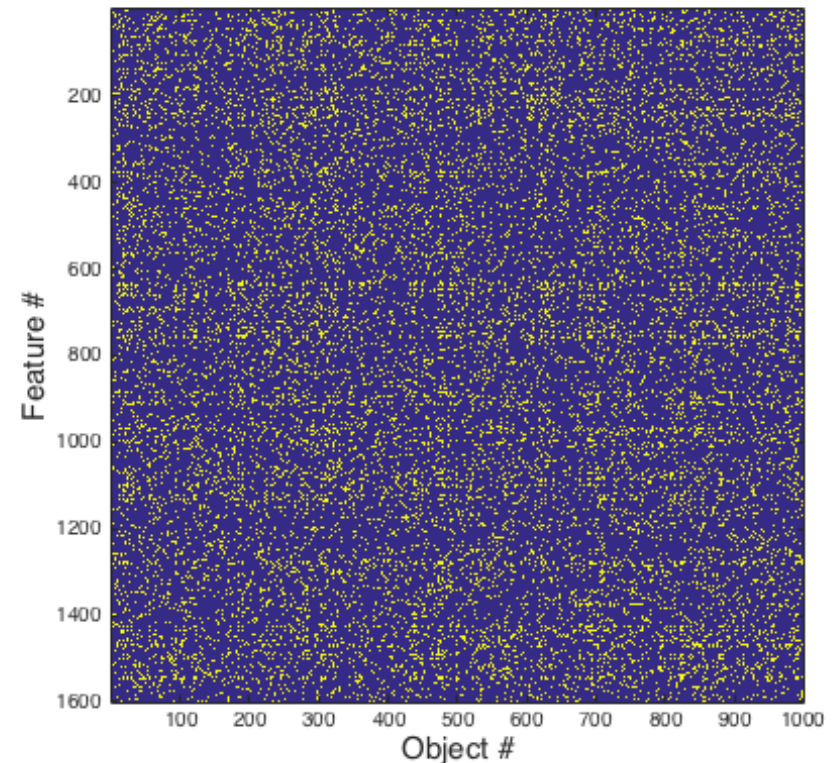
Some categories make more “sense”, or are more “coherent” than others.

i.e. “incoherent” = the set of all things that are blue
i.e. “coherent” = the set of all things that are dogs

What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

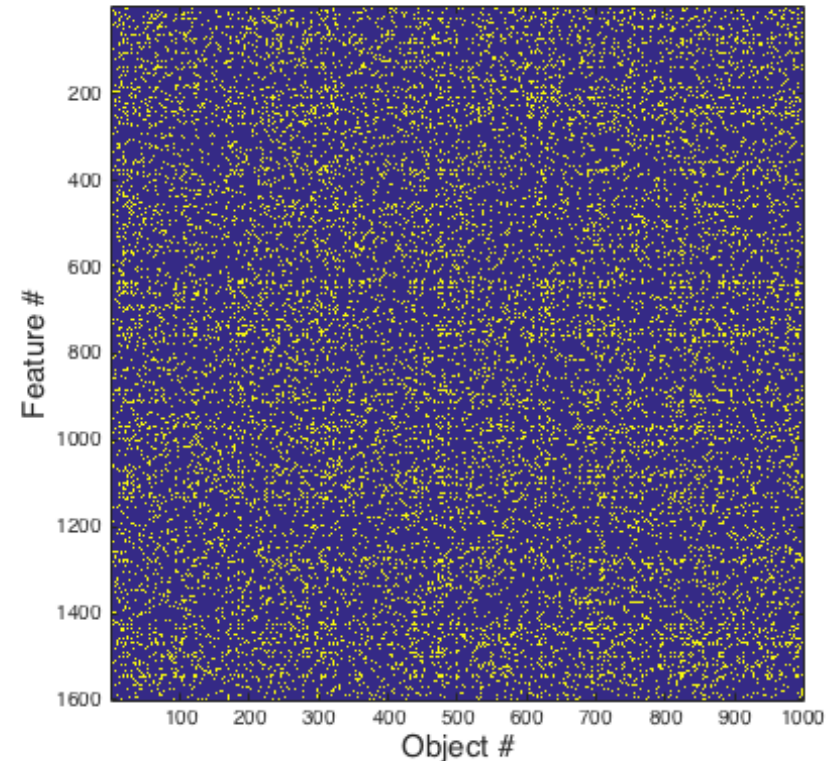
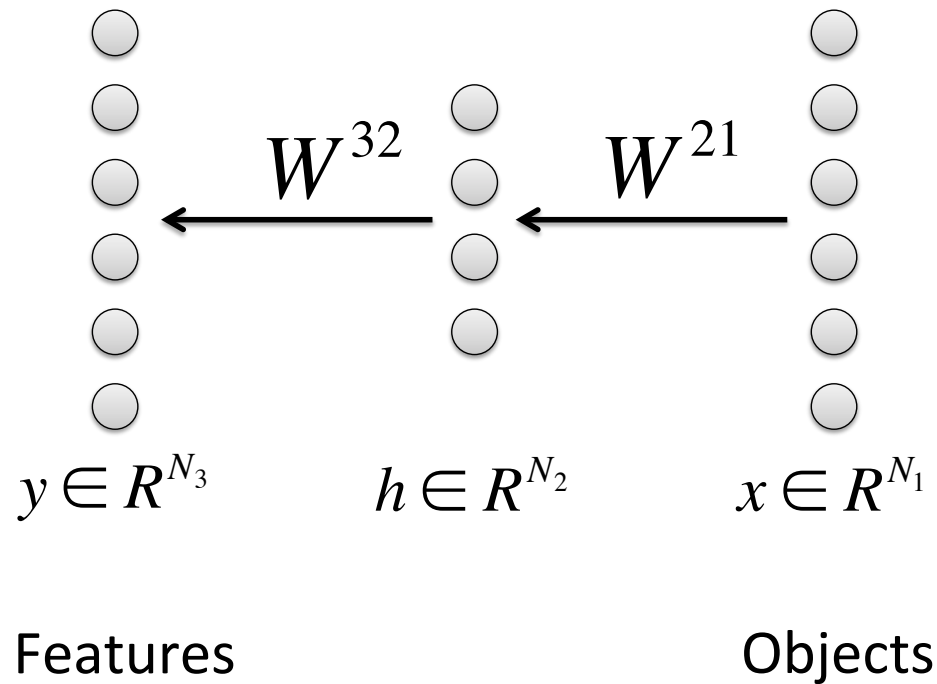
A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?



What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

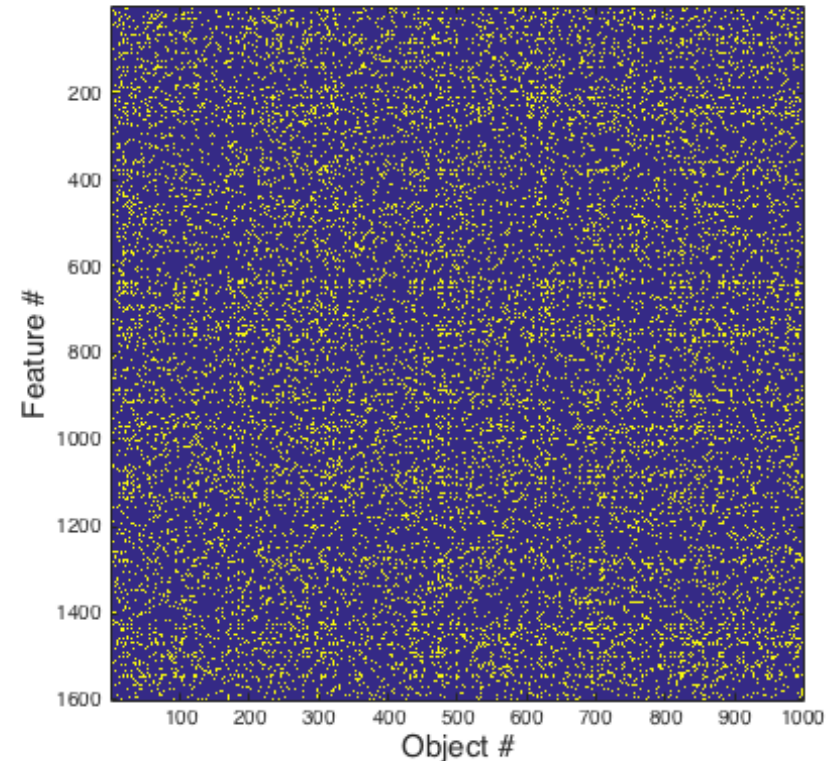
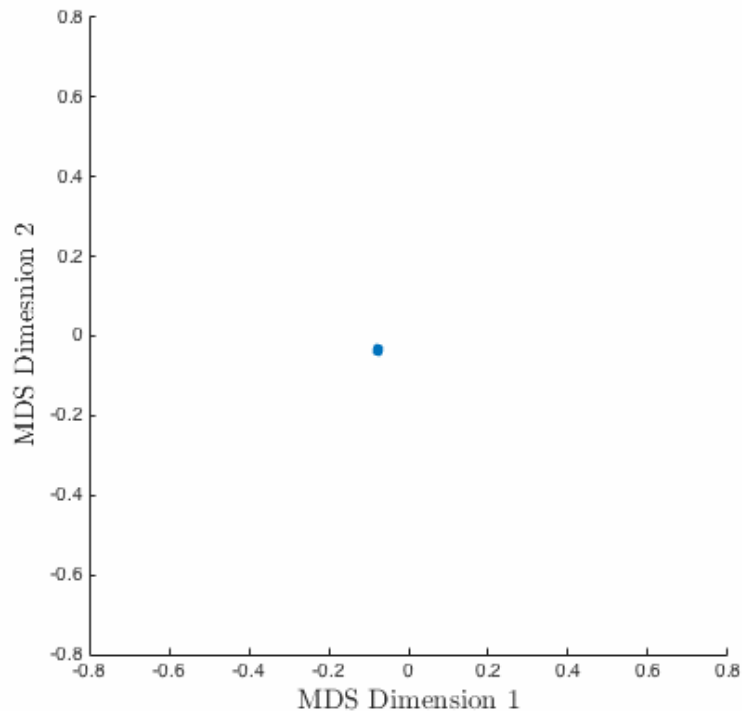
A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?



What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

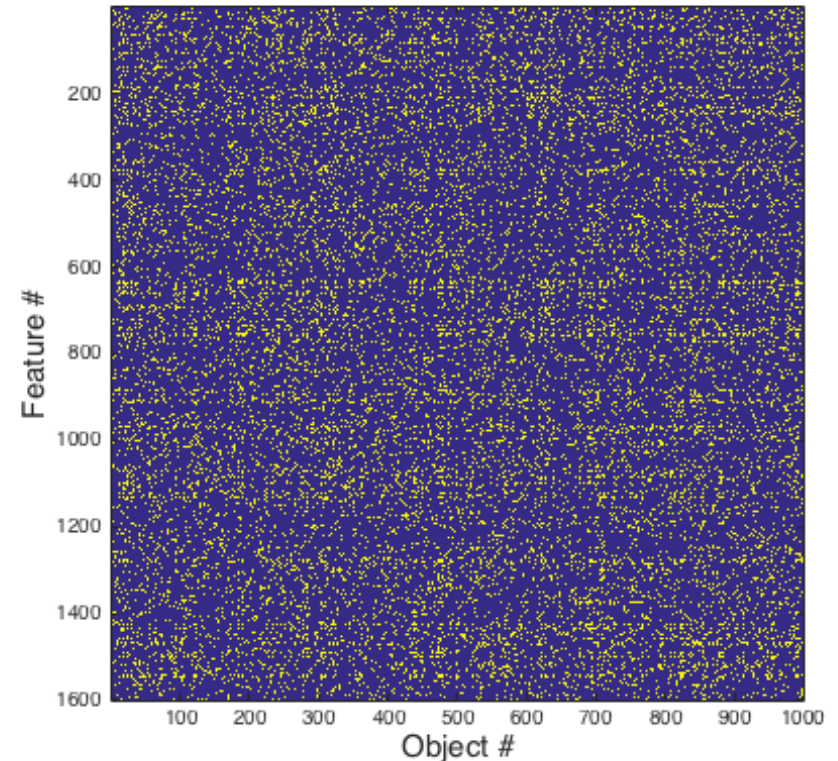
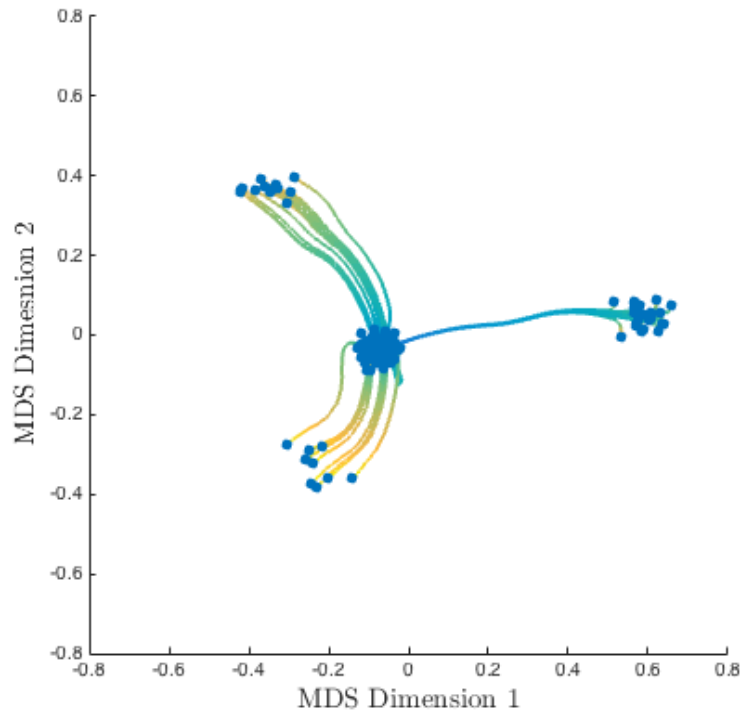
A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?



What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

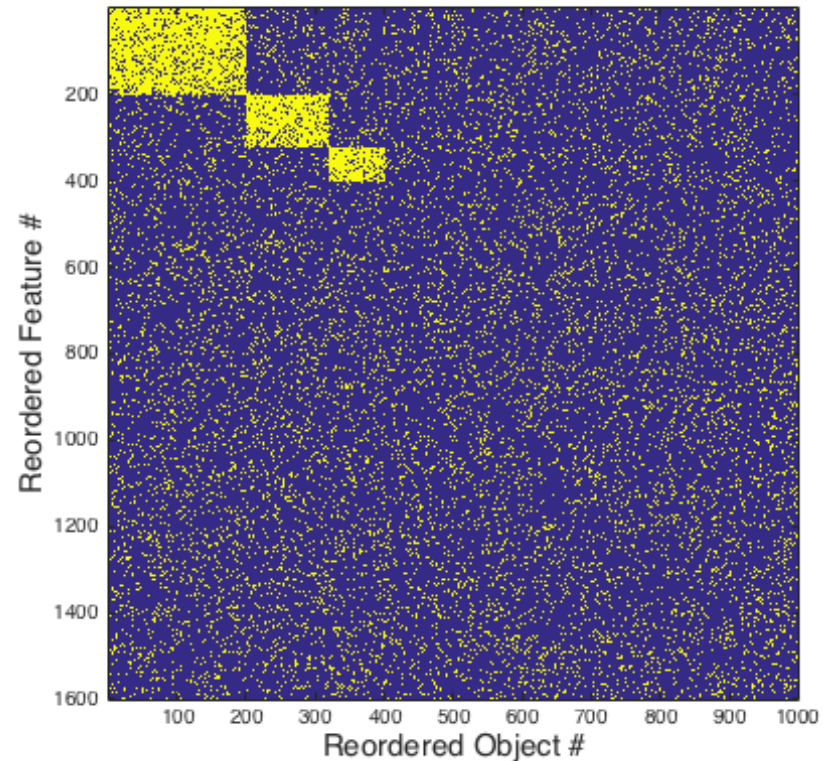
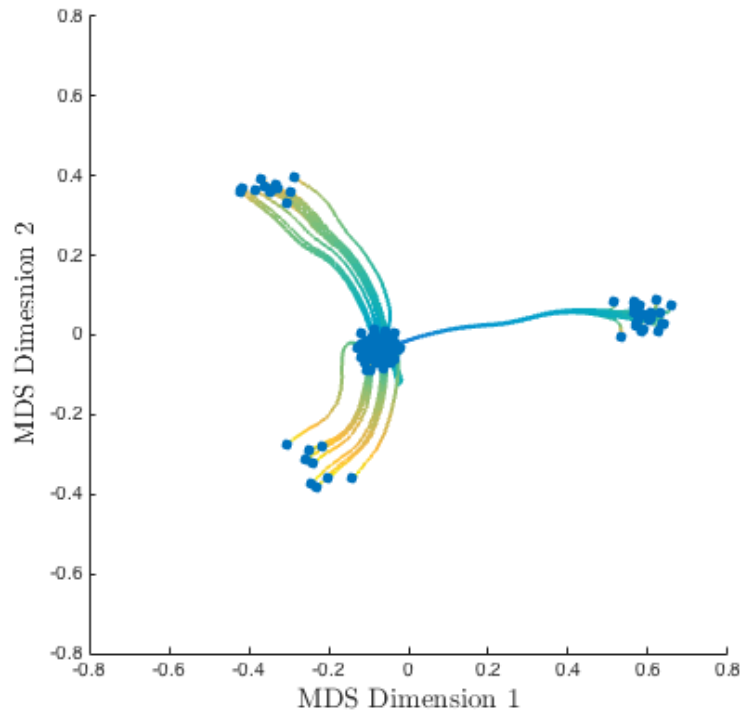
A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?



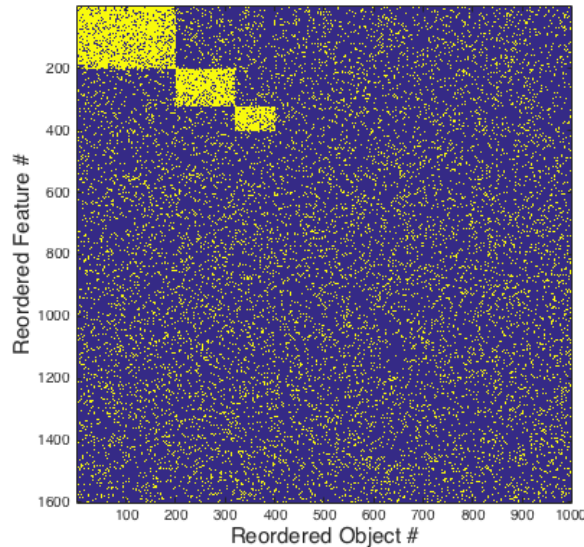
What is a category and what makes it “coherent?”

A simple proposal: A category is a subset of objects sharing a subset of features important for that category.

A conceptual Gordian knot bedeviling the field of category learning in psychology: How does one learn a category?



What is a category and what makes it “coherent?”



Toy model for statistical structure of the world:

N_o = Total number of objects

N_f = Total number of features

K_o = Number of objects in a category

K_f = Number of features important

If an object is in a category and a feature is important for that category, then the probability this object has that feature is p .

Otherwise, the probability any other object has any feature is $q < p$.

For what values of N_o , K_o , N_f , K_f , p and q can a category be learned?

$$\frac{p - q}{\sqrt{q(1 - q)}} K_o K_f \geq \sqrt{N_o N_f}$$

How fast can it be learned?

↑
Learning time is inversely related.

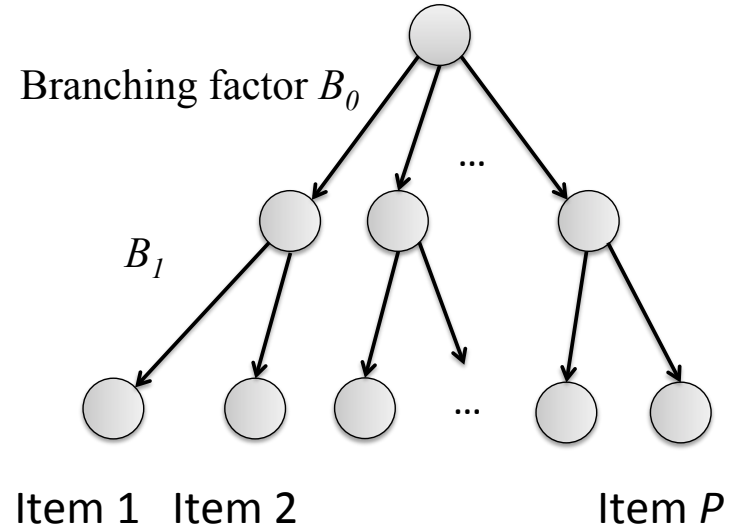
What is a category and what makes it “coherent?”

i.e. “incoherent” = the set of all things that are blue

i.e. “coherent” = the set of all things that are dogs

A natural definition category coherence is the singular value associated with object analyzers and feature synthesizers

For hierarchically structured data:



Coherence = similarity of descendants – similarity to nearest out-category

Mathematical Theorem: Coherent categories are learned faster!

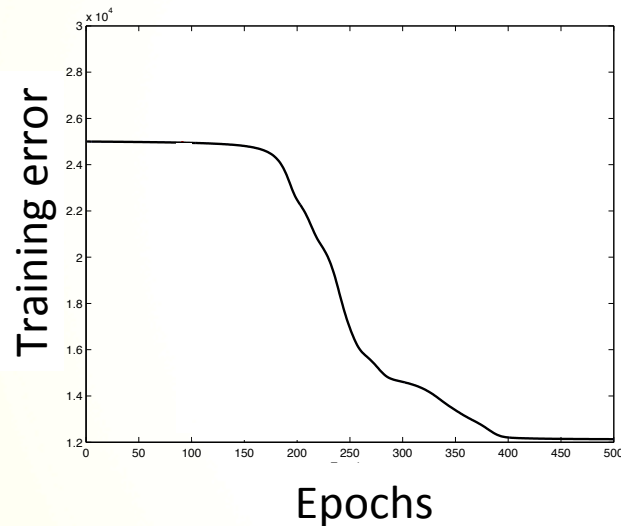
The category coherence of any one category is an emergent property of the entire statistical structure of the world: in particular the structure of individual categories and their relations to each other!

Towards a theory of deep learning dynamics

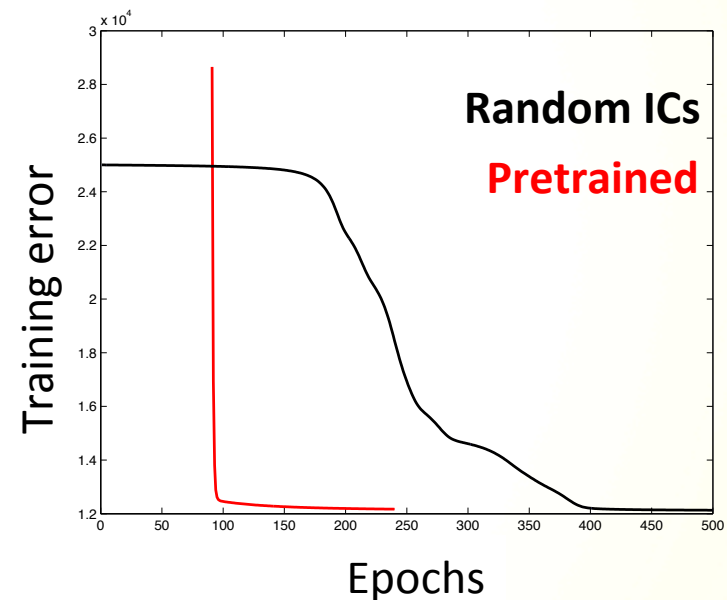
- The dynamics of learning in deep networks is non-trivial – i.e. plateaus and sudden transitions to better performance
- How does training time scale with depth?
- How should the learning rate scale with depth?
- How do different weight initializations impact learning speed?
- We will find that weight initializations with *critical dynamics* can aid deep learning and generalization.

Nontrivial learning dynamics

Plateaus and sudden transitions



Faster convergence from pretrained initial conditions



- Build intuitions for nonlinear case by analyzing linear case

Deeper networks

- Can generalize to arbitrary depth network
- Each effective singular value a evolves independently

$$\tau \frac{d}{dt} a = (N_l - 1) a^{2-2/(N_l-1)} (s - a)$$

τ	1/Learning rate
s	Singular value
N_l	# layers

- In deep networks, combined gradient is $O(N_l/\tau)$



$$a = \prod_{i=1}^{N_l-1} W_i$$

Deep linear learning speed

- Intuition (see paper for details):
 - Gradient norm $O(N_l)$
 - Learning rate $O(1/N_l)$ ($N_l = \# \text{ layers}$)
 - Learning time $O(1)$
- Deep learning *can be fast* with the right ICs.

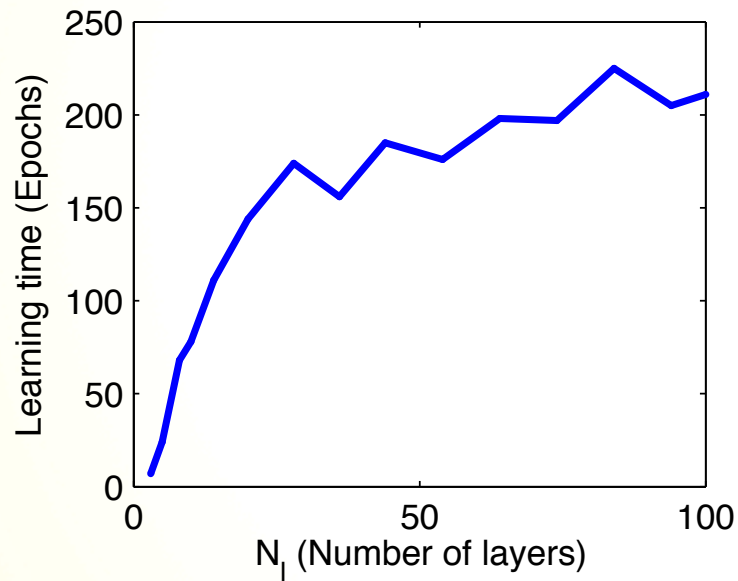
MNIST learning speeds

- Trained deep *linear* nets on MNIST
- Depths ranging from 3 to 100
- 1000 hidden units/layer (overcomplete)
- Decoupled initial conditions with fixed initial mode strength
- Batch gradient descent on squared error
- Optimized learning rates for each depth
- Calculated epoch at which error falls below fixed threshold



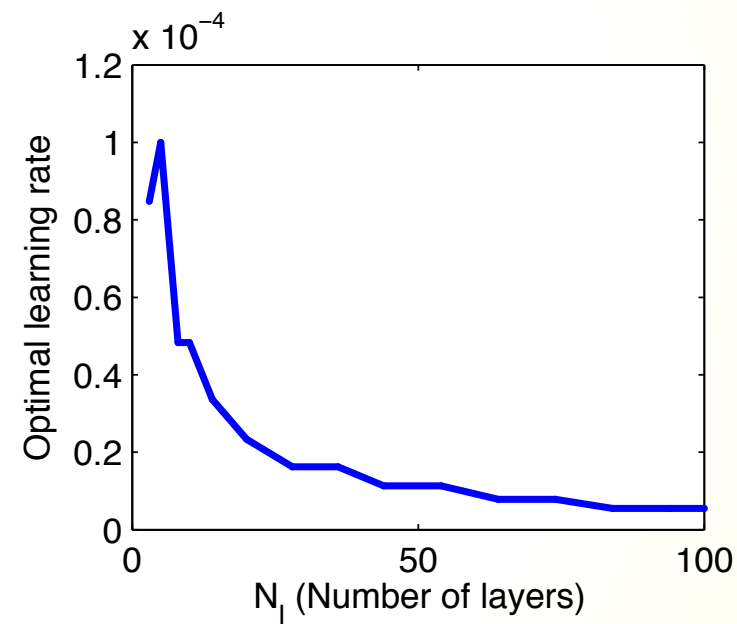
MNIST depth dependence

Time to criterion



Depth

Optimal learning rate



Depth

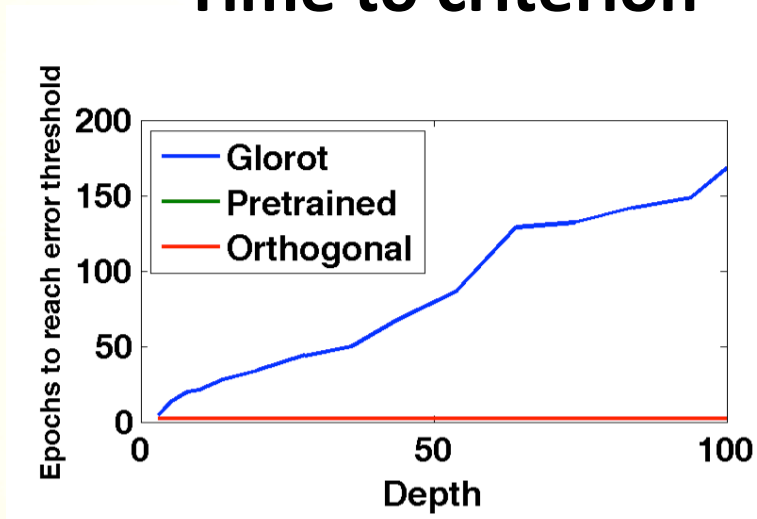
Deep linear networks

- Deep learning *can be fast* with decoupled ICs and $O(1)$ initial mode strength.
How to find these?
- Answer: Pre-training and random orthogonal initializations can find these special initial conditions that allow depth independent training times!!
- But scaled random Gaussian initial conditions on weights cannot.

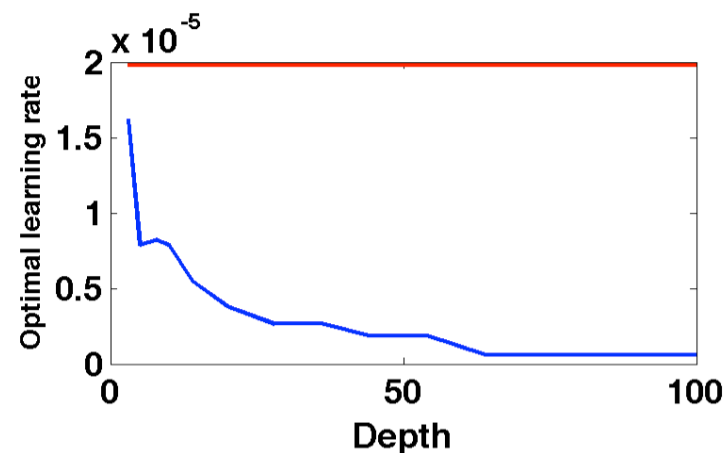
Depth-independent training time

- Deep *linear* networks on MNIST
- Scaled random Gaussian initialization (Glorot & Bengio, 2010)

Time to criterion



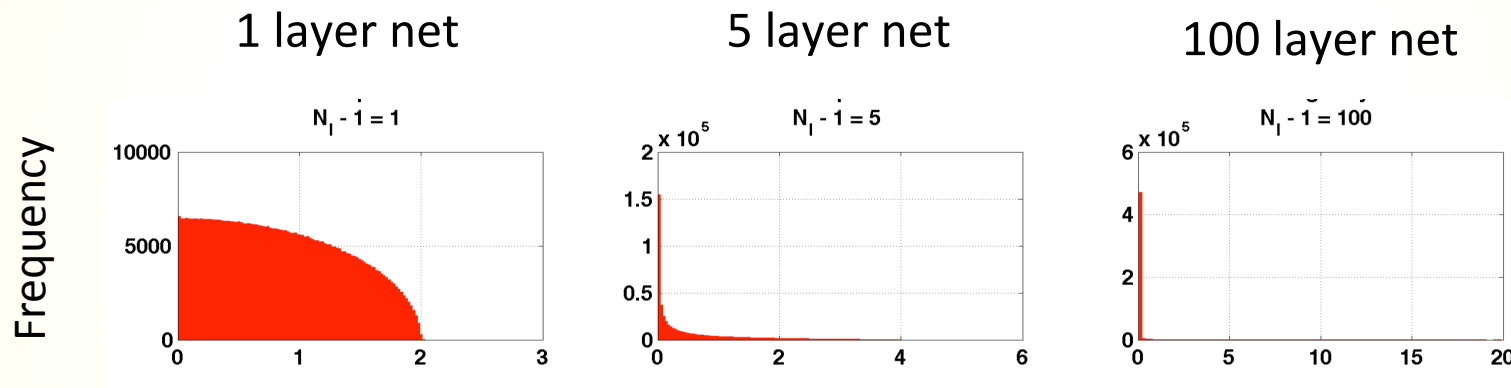
Optimal learning rate



- Pretrained and orthogonal have fast **depth-independent** training times!

Random vs orthogonal

- Gaussian preserves norm of random vector *on average*

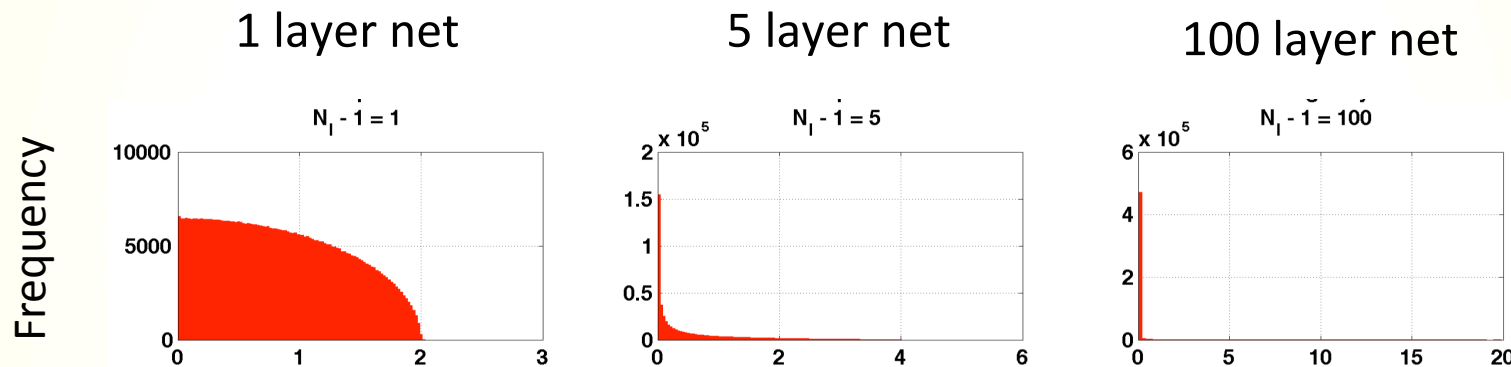


Singular values of $W^{tot} = \prod_{i=1}^{N_l-1} W^i$

- *Attenuates* on subspace of high dimension
- *Amplifies* on subspace of low dimension

Random vs orthogonal

- Glorot preserves norm of random vector *on average*



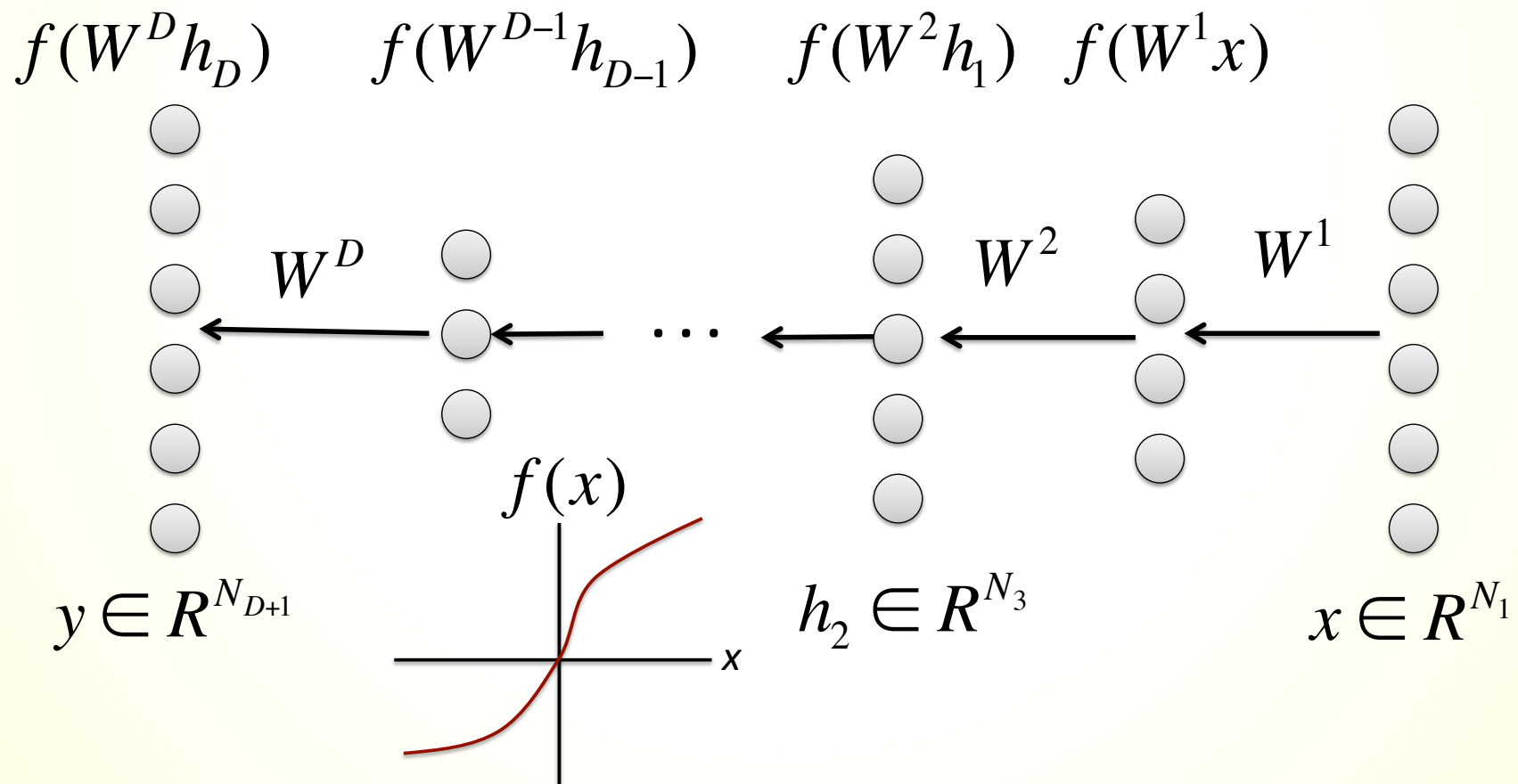
Singular values of $W^{tot} = \prod_{i=1}^{N_l-1} W^i$

- Orthogonal preserves norm of all vectors *exactly*

All singular values of $W^{tot} = 1$

Deeper network learning dynamics

- Jacobian that back-propagates gradients can explode or decay

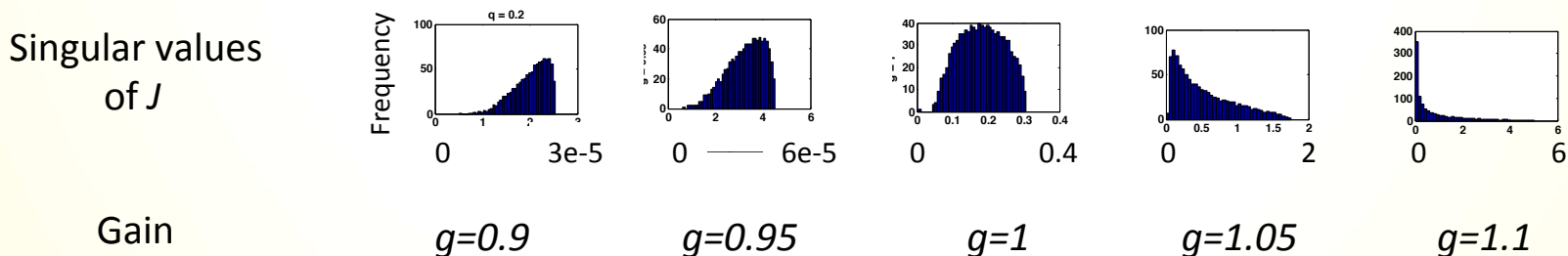


Extensive Criticality yields Dynamical Isometry in *nonlinear* nets

Suggests initialization for *nonlinear* nets

- near-isometry on subspace of large dimension
- Singular values of *end-to-end* Jacobian $J_{ij}^{N_l,1}(x^{N_l}) \equiv \frac{\partial x_i^{N_l}}{\partial x_j^1} \Big|_{x^{N_l}}$ concentrated around 1.

Scale orthogonal matrices by gain g to counteract contractive nonlinearity



Just beyond *edge of chaos* ($g>1$) may be good initialization

Dynamic Isometry Initialization

- $g > 1$ speeds up **30 layer nonlinear** nets
 - Tanh network, softmax output, 500 units/layer
 - No regularization (weight decay, sparsity, dropout, etc)

MNIST Classification error, epoch 1500	Train Error (%)	Test Error (%)
Gaussian (g=1, random)	2.3	3.4
g=1.1, random	1.5	3.0
g=1, orthogonal	2.8	3.5
Dynamic Isometry (g=1.1, orthogonal)	0.095	2.1

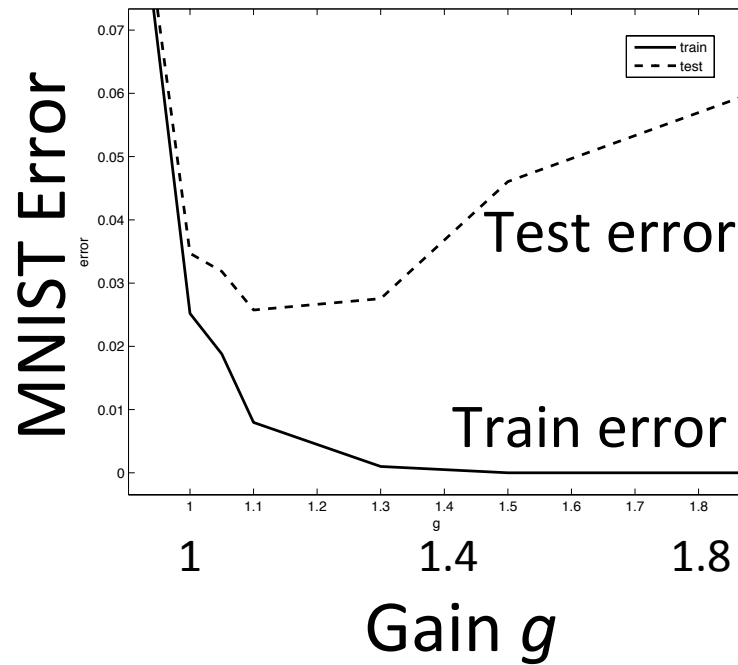
- Dynamic isometry reduces test error by 1.4% pts

Summary

- Deep linear nets have **nontrivial nonlinear learning dynamics**.
- Learning time inversely proportional to strength of input-output correlations.
- With the right initial weight conditions, number of training epochs can remain finite as depth increases.
- Dynamically critical networks just beyond the edge of chaos enjoy **depth-independent** learning times.

Beyond learning: criticality and generalization

- Deep networks + large gain factor g train exceptionally quickly
- But large g incurs heavy cost in generalization performance



- Suggests small initial weights regularize towards smoother functions

Some of the theoretical puzzles of deep learning

Trainability: if a good network solution exists with small training error, how do we find it? And what makes a learning problem difficult?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks ICLR 2014.

A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, CogSci 2013.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

Generalizability: what principles do deep networks use to place probability / make decisions in regions of input space with little data?

M. Advani and S. Ganguli, Statistical Mechanics of Optimal Convex Inference in High Dimensions, Physical Review X, 2016.

Expressiveness, Memorization, Stability, and Flat versus sharp minima.

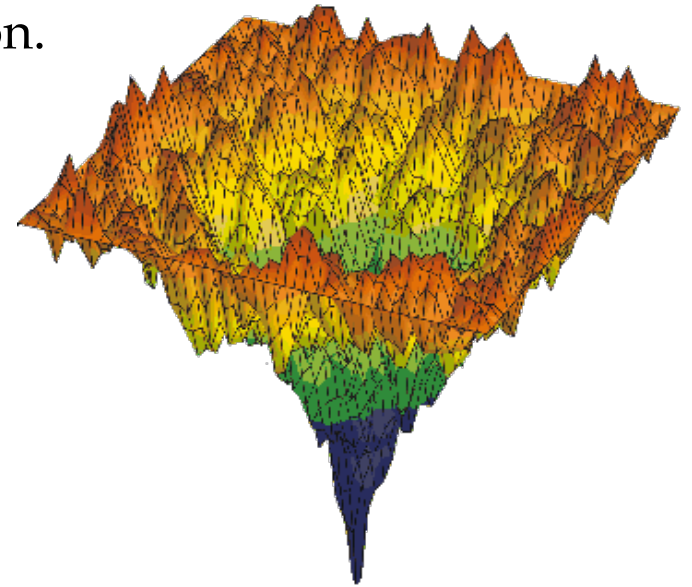
High dimensional nonconvex optimization

It is often thought that local minima at high error stand as a major impediment to non-convex optimization.

In random non-convex error surfaces over high dimensional spaces, local minima at high error are exponentially rare in the dimensionality.

Instead saddle points proliferate.

We developed an algorithm that rapidly escapes saddle points in high dimensional spaces.



Identifying and attacking the saddle point problem in high dimensional non-convex optimization.
Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio. NIPS 2014

A. Choromanska, M. B. Henaff, M. Mathieu, G. Ben Arous, Y. LeCun, The Loss Surfaces of Multilayer Networks, in the International Conference on Artificial Intelligence and Statistics (AISTATS), 2015 pdf

General properties of error landscapes in high dimensions

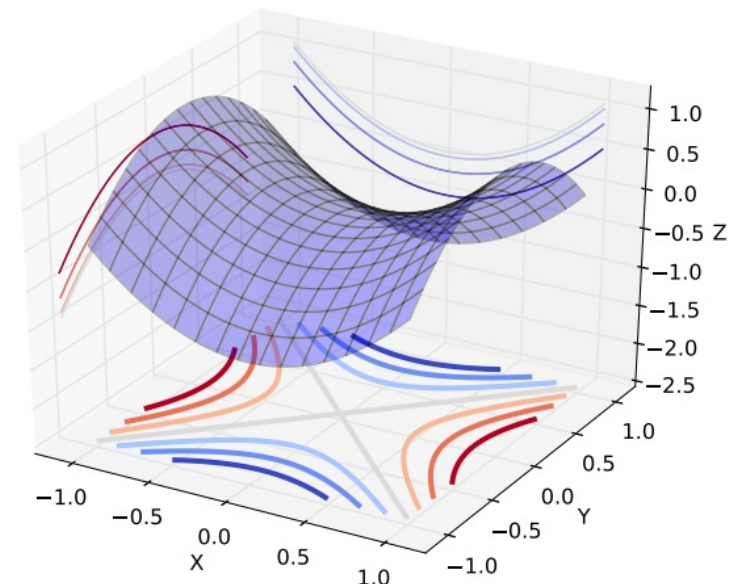
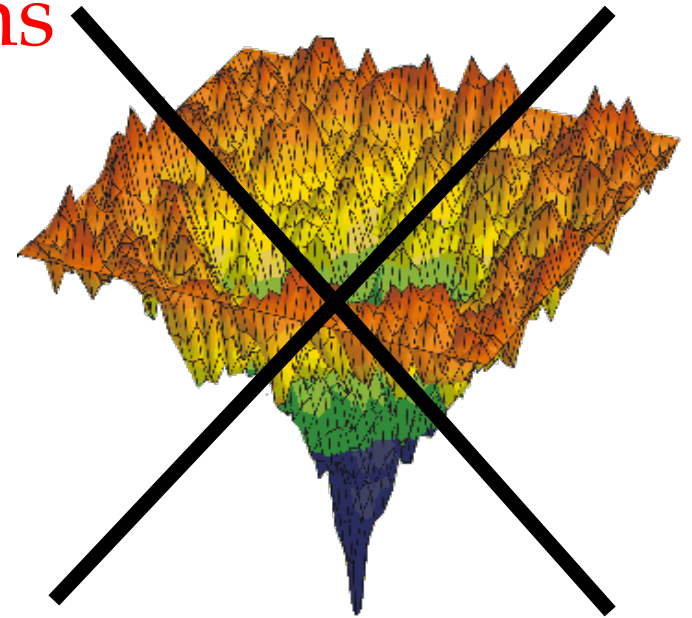
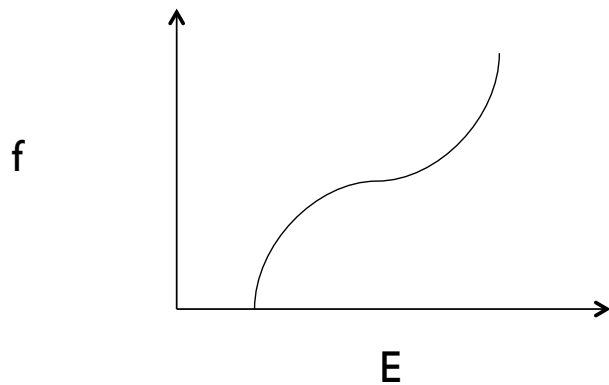
From statistical physics:

Consider a random Gaussian error landscape over N variables.

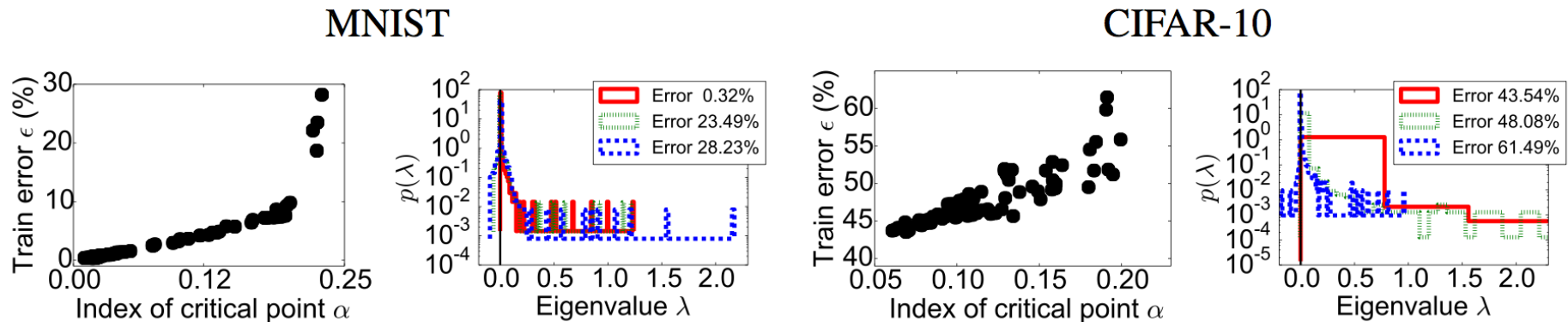
Let x be a critical point.

Let E be its error level.

Let f be the fraction of negative curvature directions.

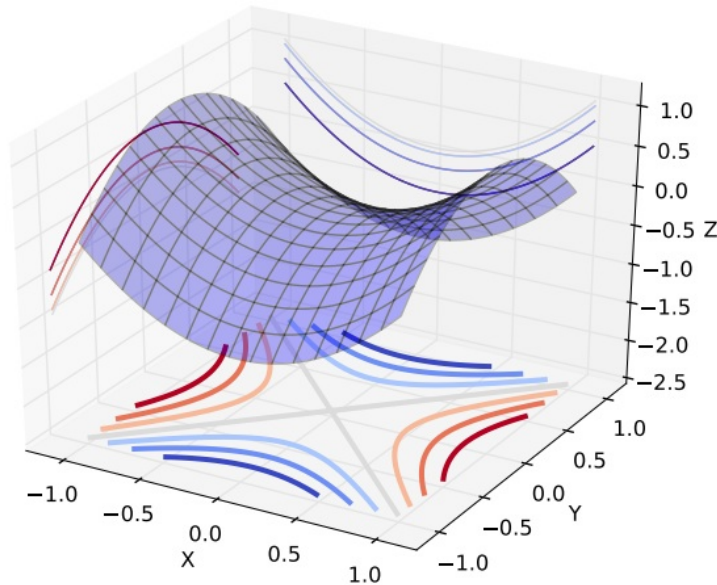


Properties of Error Landscapes on the Synaptic Weight Space of a Deep Neural Net



Qualitatively consistent with the statistical physics theory of random error landscapes

How to descend saddle points



Newton's Method

$$\Delta x = -H^{-1} \nabla f(x)$$

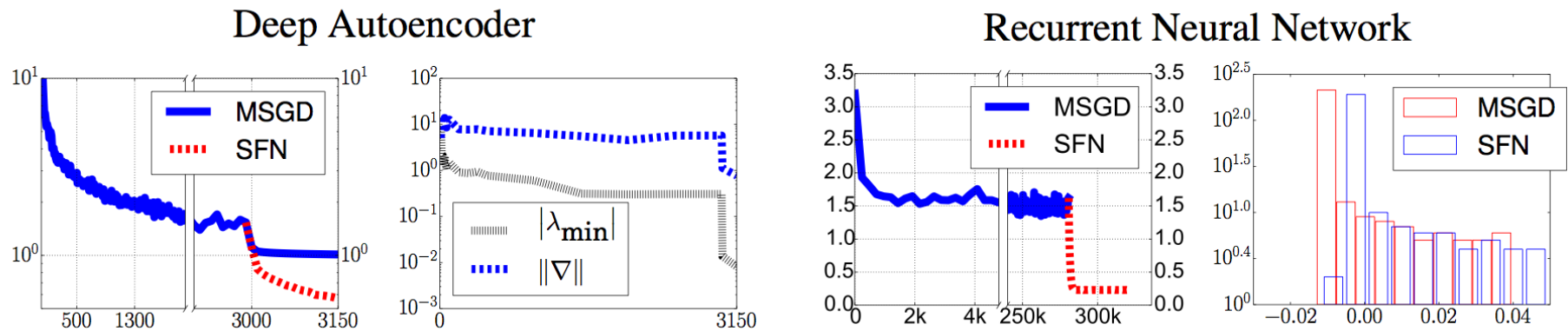
Saddle Free Newton's Method

$$\Delta x = -|H|^{-1} \nabla f(x)$$

Intuition: saddle points **attract** Newton's method, but **repel** saddle free Newton's method.

Derivation: minimize a **linear** approximation to $f(x)$ within a trust region in which the linear and quadratic approximations agree

Performance of saddle free Newton in learning deep neural networks.



When stochastic gradient descent appears to plateau, switching to saddle Free newton escapes the plateau.

Some of the theoretical puzzles of deep learning

Trainability: if a good network solution exists with small training error, how do we find it? And what makes a learning problem difficult?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks ICLR 2014.

A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, CogSci 2013.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, under review, NIPS 2016.

Generalizability: what principles do deep networks use to place probability / make decisions in regions of input space with little data?

M. Advani and S. Ganguli, Statistical Mechanics of Optimal Convex Inference in High Dimensions, Physical Review X, 2016.

Expressiveness, Memorization, Stability, and Flat versus sharp minima.

A theory of deep neural expressivity through transient chaos

Stanford



Ben Poole

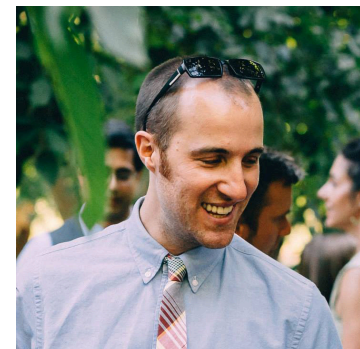


Subhaneil
Lahiri



Maithra
Raghu

Google



Jascha
Sohl-Dickstein

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

On the expressive power of deep neural networks, M. Raghu, B. Poole, J. Kleinberg, J. Sohl-Dickstein, S. Ganguli, under review, ICML 2017.

Seminal works on the expressive power of depth

Networks with one hidden layer are universal function approximators.

So why do we need depth?

Universal function approximation theorems yield no guarantees on the size of the hidden layer needed to approximate a function well.

Overall idea: there exist certain (special?) functions that can be computed:

- a) efficiently using a deep network (poly # of neurons in input dimension)
- b) but not by a shallow network (requires exponential # of neurons)

Intellectual traditions in boolean circuit theory: parity function is such a function for boolean circuits.

Seminal works on the expressive power of depth

Nonlinearity

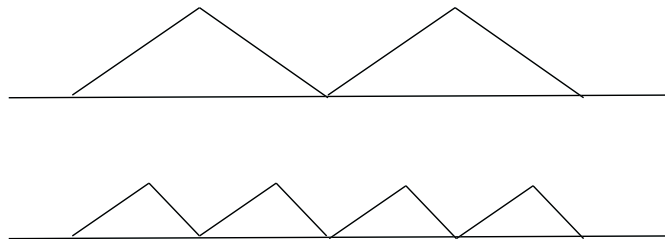
Rectified Linear Unit (ReLU)

Measure of Functional Complexity

Number of linear regions

There exists a function computable by a deep network where the number of linear regions is exponential in the depth.

To approximate this function with a shallow network, one would require exponentially many more neurons.



Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio.
On the number of linear regions of deep neural networks, NIPS 2014

Seminal works on the expressive power of depth

Nonlinearity

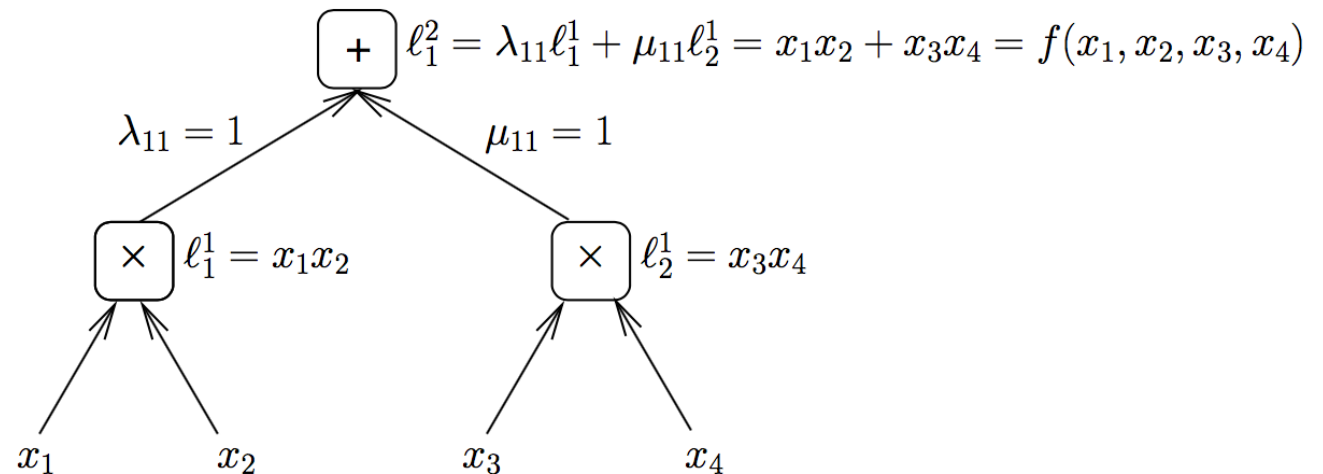
Measure of Functional Complexity

Sum-product network

Number of monomials

There exists a function computable by a deep network where the number of unique monomials is exponential in the depth.

To approximate this function with a shallow network, one would require exponentially many more neurons.



Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks, NIPS 2011.

Questions

How natural are these functions from the perspective of AI?

Are such functions rare curiosities?

Or is this phenomenon much more generic than these specific examples?

In some sense, is any function computed by a generic deep network not efficiently computable by a shallow network?

If so we would like a theory of deep neural expressivity that demonstrates this for

- 1) Arbitrary nonlinearities
- 2) A natural, general measure of functional complexity.

Limitations of prior work

Theoretical technique

Nonlinearity

Measure of Functional
Complexity

Combinatorics/
Hyperplane Arrangements

ReLU

Number of linear regions

Polynomial expansion

Sum-product

Number of monomials

Algebraic topology

Pfaffian

Sum of betti numbers

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *Neural Networks and Learning Systems, IEEE Transactions on*, 2014.

**Riemannian geometry +
Dynamical mean field theory**

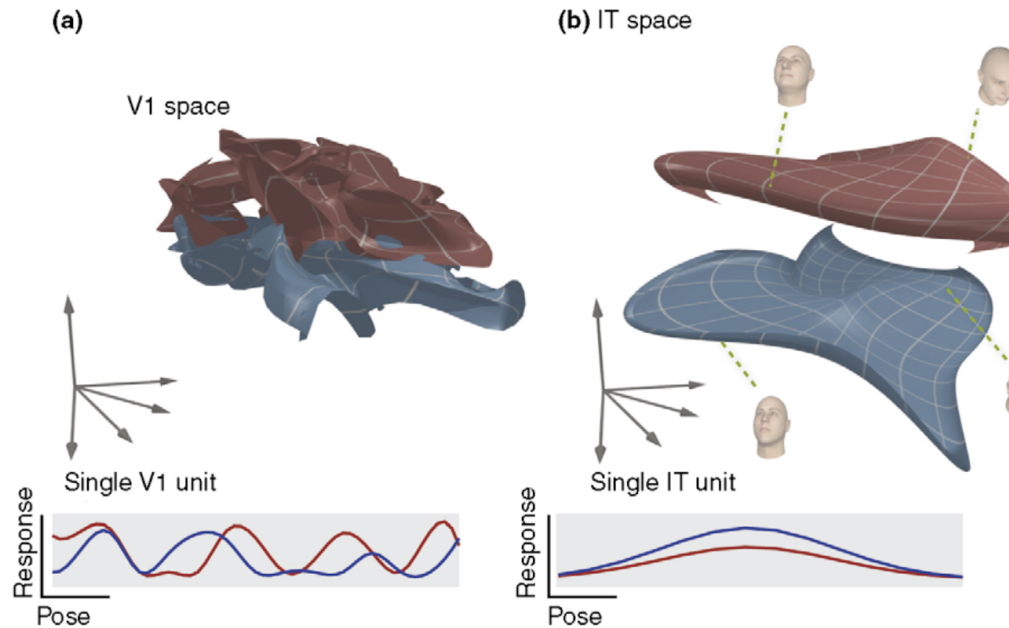
Arbitrary

**Extrinsic
Curvature**

We will show that even in generic, random deep neural networks, measures of functional curvature grow exponentially with depth but not width!

More over the origins of this exponential growth can be traced to chaos theory.

Another perspective on the advantage of depth: disentangling

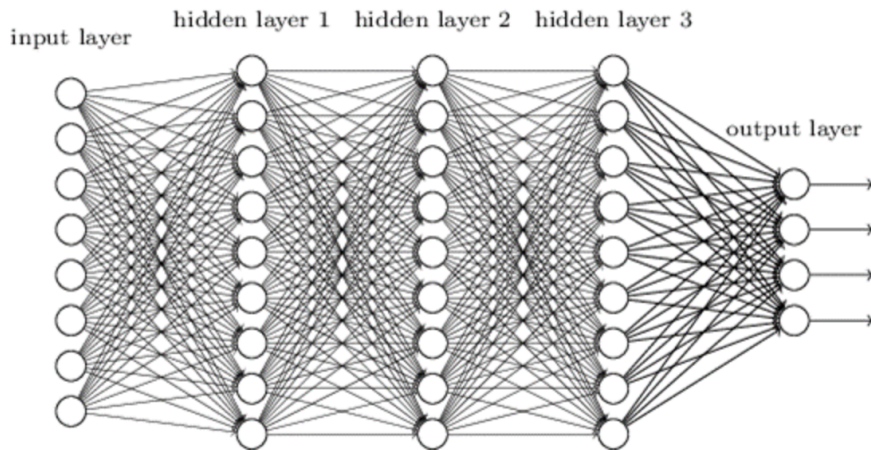


How can we mathematically formalize the notion of disentangling in deep networks?

How do we use this mathematical formalization to quantitatively assess the disentangling power of deep versus shallow networks?

We will show that deep networks can disentangle manifolds whose curvature grows exponentially with depth!

A maximum entropy ensemble of deep random networks



N_l = number of neurons in layer l

D = depth ($l = 1, \dots, D$)

$\mathbf{x}^l = \phi(\mathbf{h}^l)$

$\mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l$

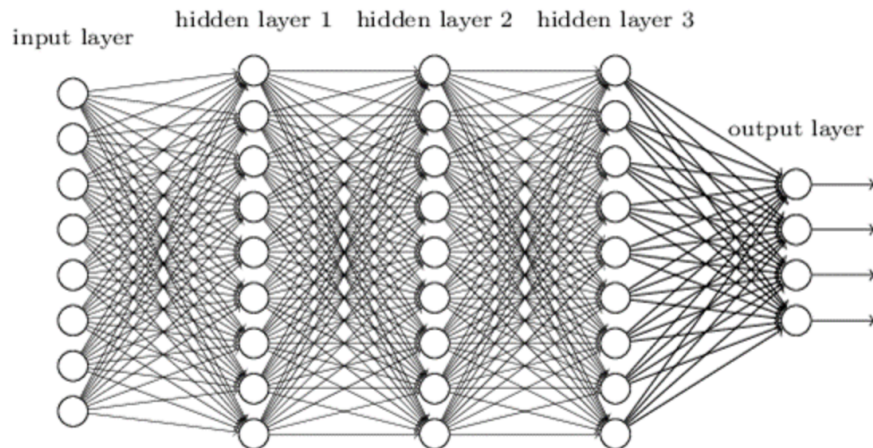
Structure:

i.i.d. random Gaussian weights and biases:

$$\mathbf{W}_{ij}^l \leftarrow \mathcal{N}\left(0, \frac{\sigma_w^2}{N^{l-1}}\right)$$

$$\mathbf{b}_i^l \leftarrow \mathcal{N}(0, \sigma_b^2)$$

Emergent, deterministic signal propagation in random neural networks



N_l = number of neurons in layer l

D = depth ($l = 1, \dots, D$)

$\mathbf{x}^l = \phi(\mathbf{h}^l)$

$\mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l$

Question: how do simple input manifolds propagate through the layers?

A single point:

When does its length grow or shrink and how fast?

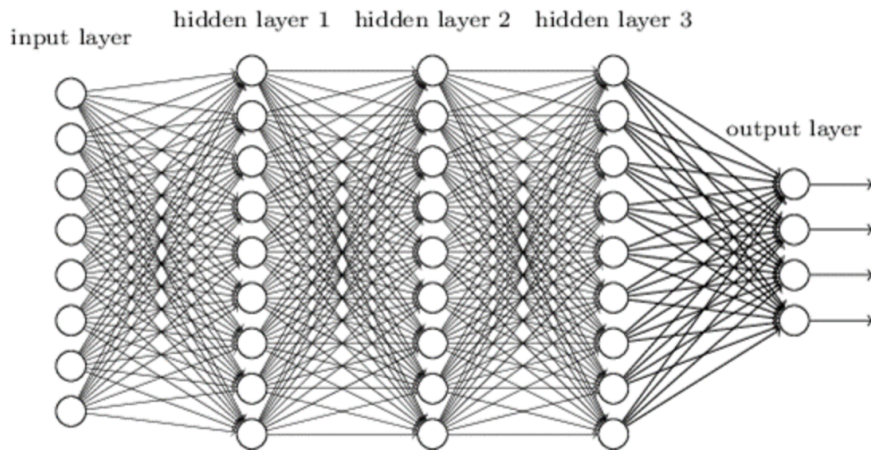
A pair of points:

Do they become more similar or more different, and how fast?

A smooth manifold:

How does its curvature and volume change?

Propagation of a single point through a deep network



N_l = number of neurons in layer l

D = depth ($l = 1, \dots, D$)

$$\mathbf{x}^l = \phi(\mathbf{h}^l)$$

$$\mathbf{h}^l = \mathbf{W}^l \mathbf{x}^{l-1} + \mathbf{b}^l$$

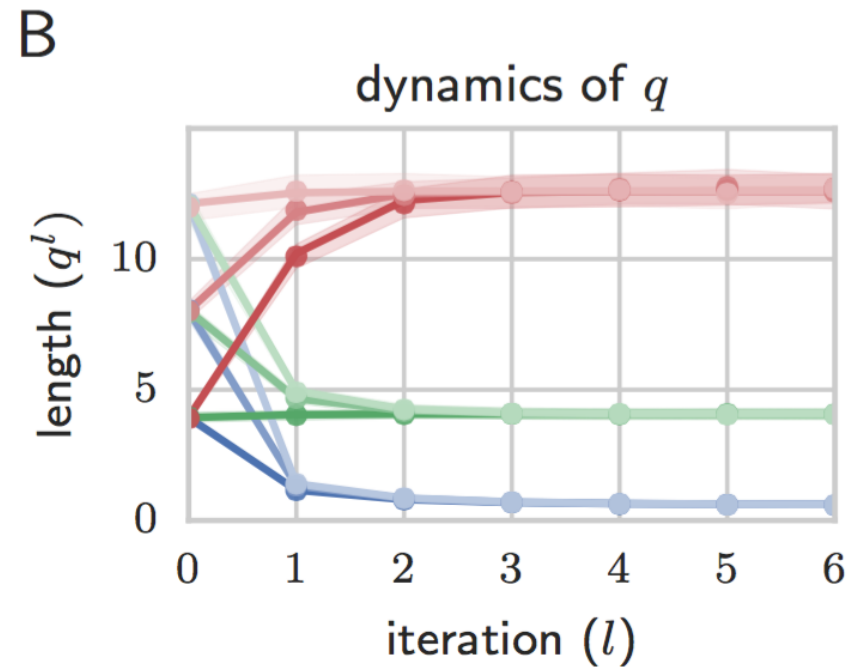
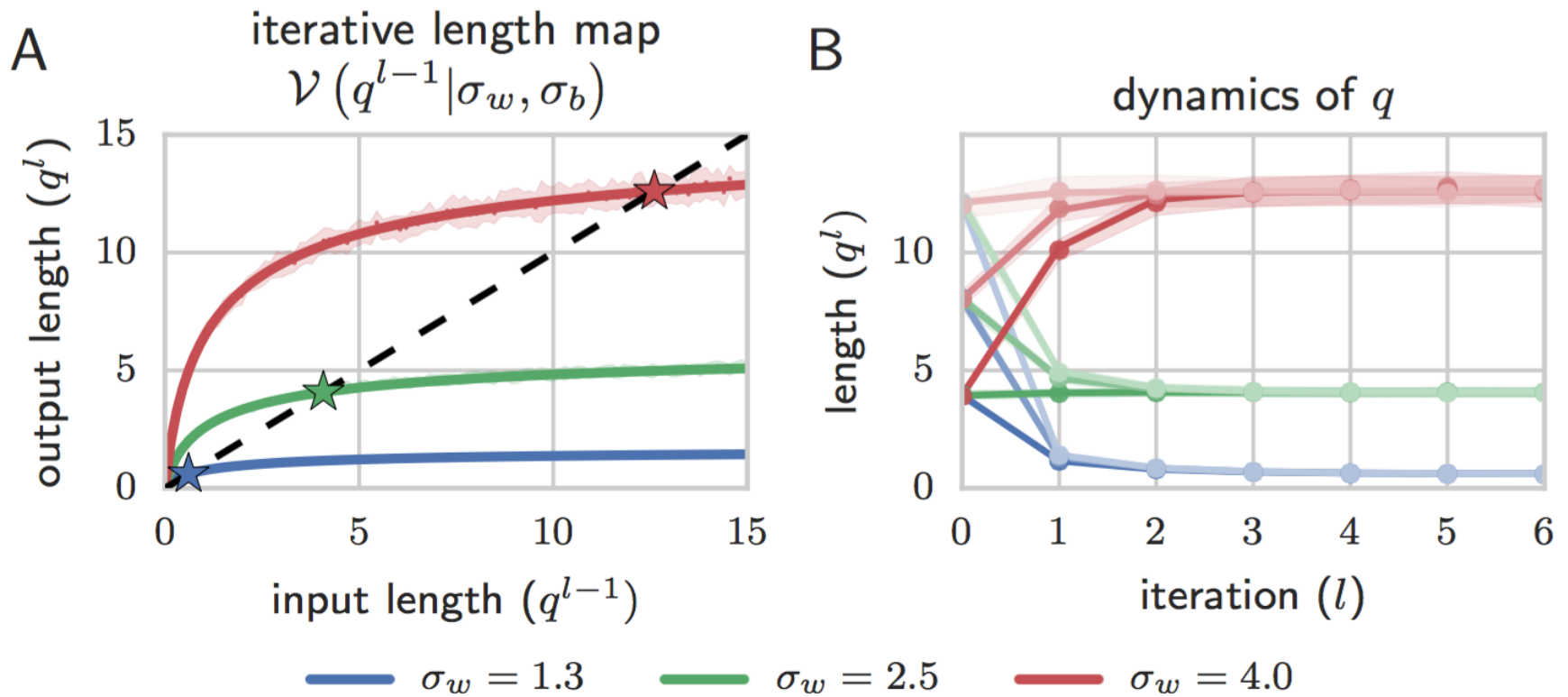
$$\mathbf{h}^l = \mathbf{W}^l \phi(\mathbf{h}^{l-1}) + \mathbf{b}^l$$

$$q^l = \frac{1}{N_l} \sum_{i=1}^{N_l} (\mathbf{h}_i^l)^2$$

$$q^l = \mathcal{V}(q^{l-1} | \sigma_w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z \phi \left(\sqrt{q^{l-1}} z \right)^2 + \sigma_b^2$$

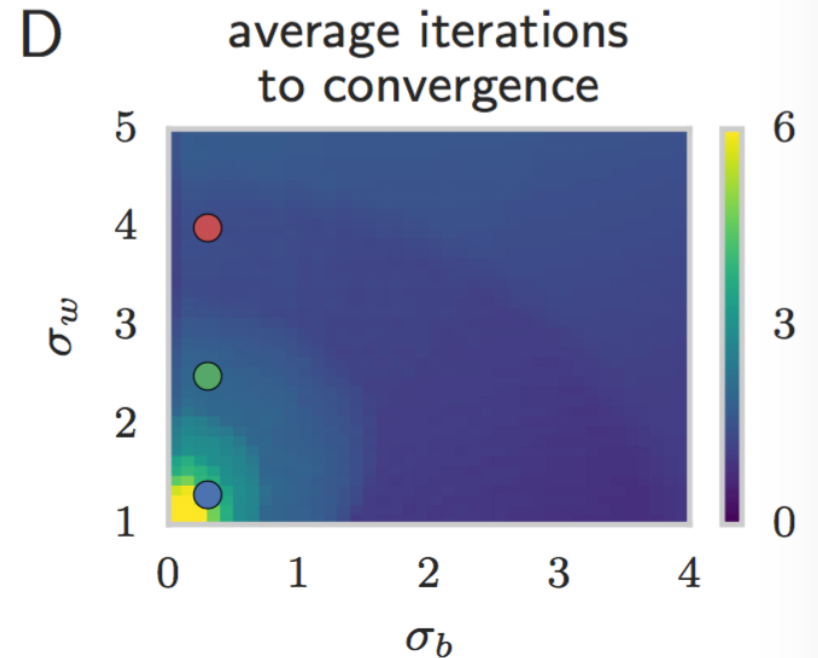
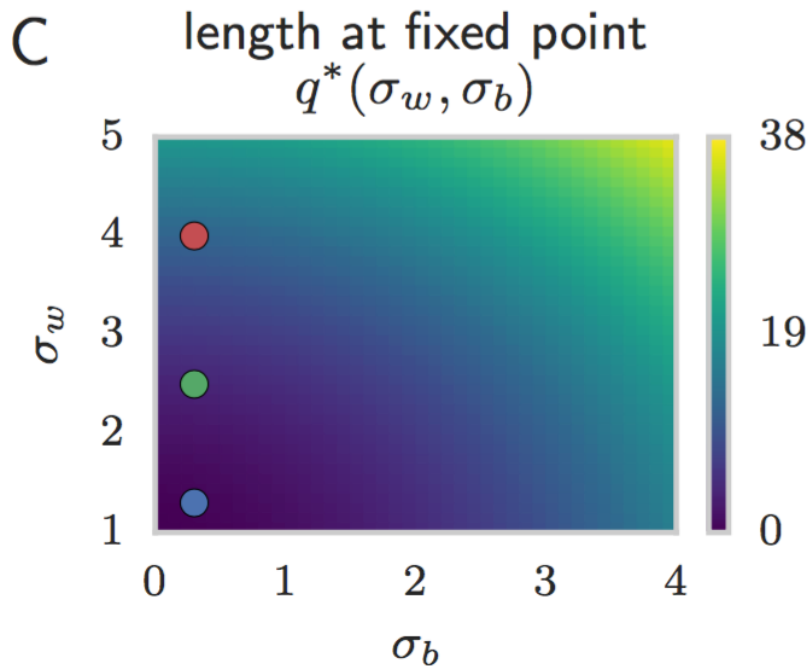
A recursion relation for the length of a point as it propagates through the network

Propagation of a single point through a deep network



$$\sigma_b = 0.3$$

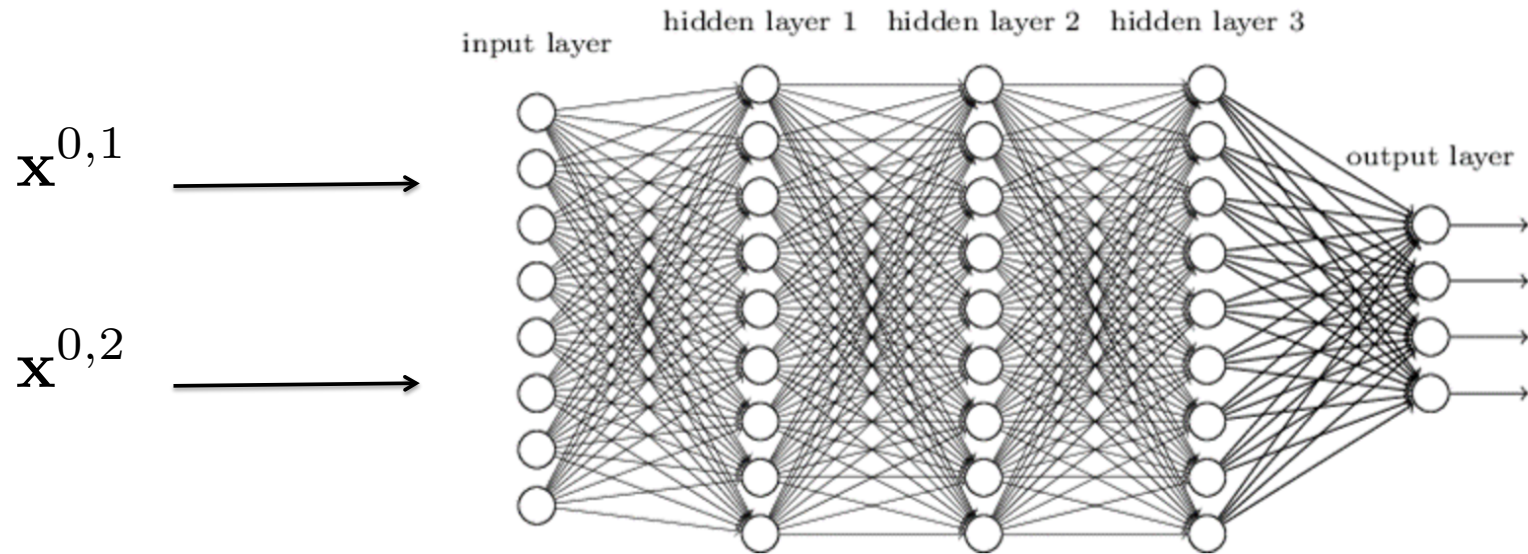
Propagation of a single point through a deep network



$$\sigma_w < 1 \quad \sigma_b = 0 : \quad q^l \rightarrow 0$$

$$\sigma_w > 1 \quad \sigma_b = 0 \quad \text{or} \quad \sigma_b \neq 0 : \quad q^l \rightarrow q^*$$

Propagation of two points through a deep network



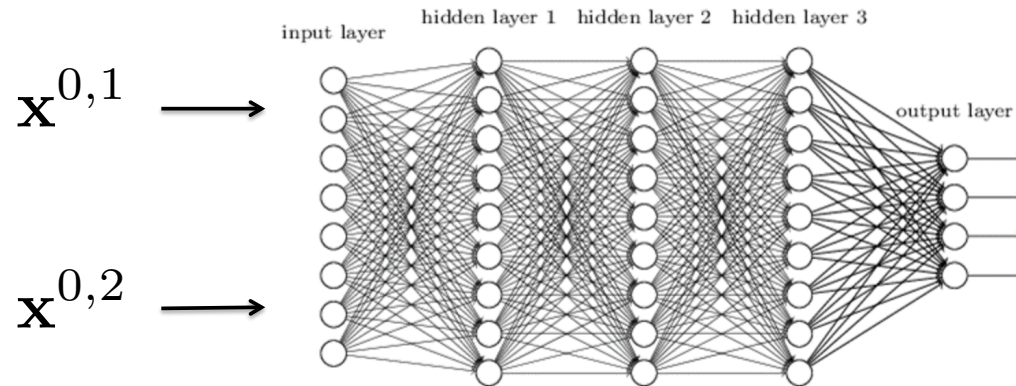
The geometry of two points in a hidden layer l is captured by the two by two matrix of inner products:

$$q_{ab}^l = \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}_i^l(\mathbf{x}^{0,a}) \mathbf{h}_i^l(\mathbf{x}^{0,b}) \quad a, b \in \{1, 2\}.$$

Of particular interest: the correlation coefficient or cosine of the angle between the two points:

$$c_{12}^l = \frac{q_{12}^l}{\sqrt{q_{11}^l} \sqrt{q_{22}^l}}$$

A theory of correlation propagation in a deep network



The geometry of two points:

$$q_{ab}^l = \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}_i^l(\mathbf{x}^{0,a}) \mathbf{h}_i^l(\mathbf{x}^{0,b}) \quad a, b \in \{1, 2\}.$$

Correlation coefficient between two points:

$$c_{12}^l = \frac{q_{12}^l}{\sqrt{q_{11}^l} \sqrt{q_{22}^l}}$$

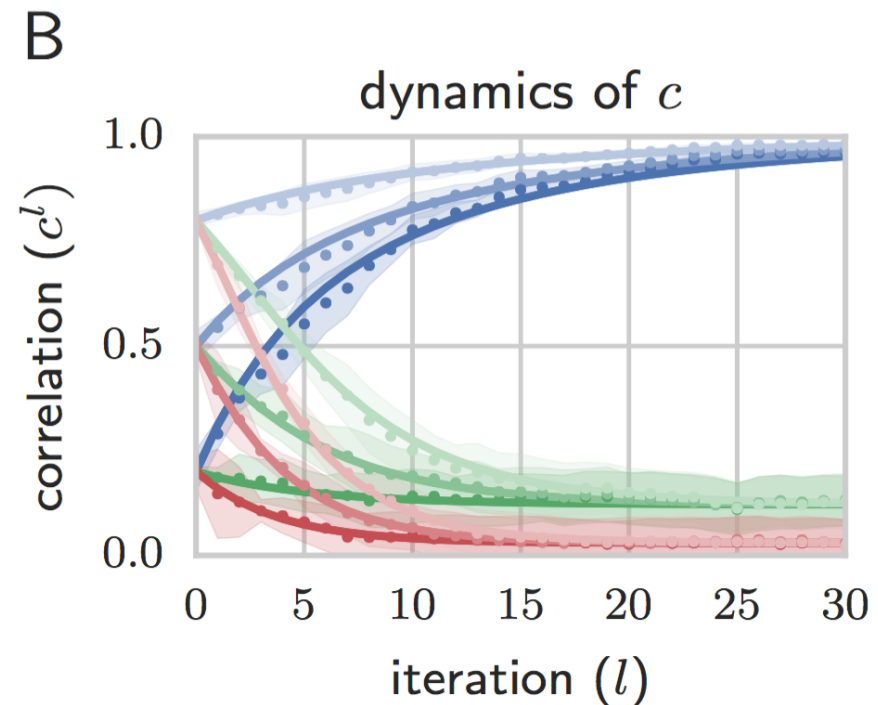
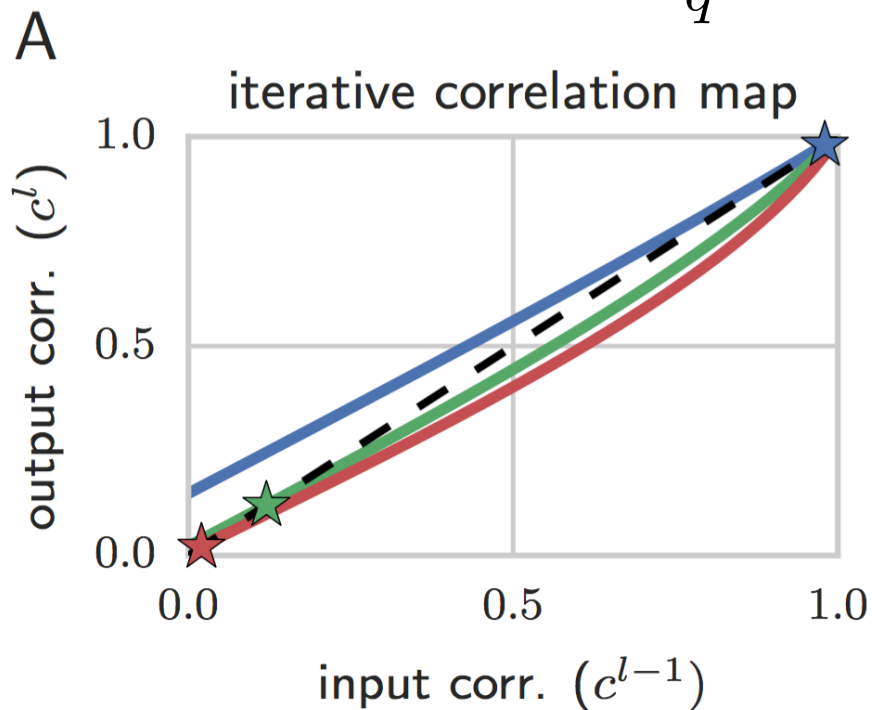
$$q_{12}^l = \mathcal{C}(c_{12}^{l-1}, q_{11}^{l-1}, q_{22}^{l-1} | \sigma_w w, \sigma_b) \equiv \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1) \phi(u_2) + \sigma_b^2,$$

$$u_1 = \sqrt{q_{11}^{l-1}} z_1, \quad u_2 = \sqrt{q_{22}^{l-1}} \left[c_{12}^{l-1} z_1 + \sqrt{1 - (c_{12}^{l-1})^2} z_2 \right],$$

A recursion relation for the correlation coeff. between two points in a deep net!

Propagation of correlations through a deep network

$$c_{12}^l = \frac{1}{q^*} \mathcal{C}(c_{12}^{l-1}, q^*, q^* | \sigma_w, \sigma_b)$$



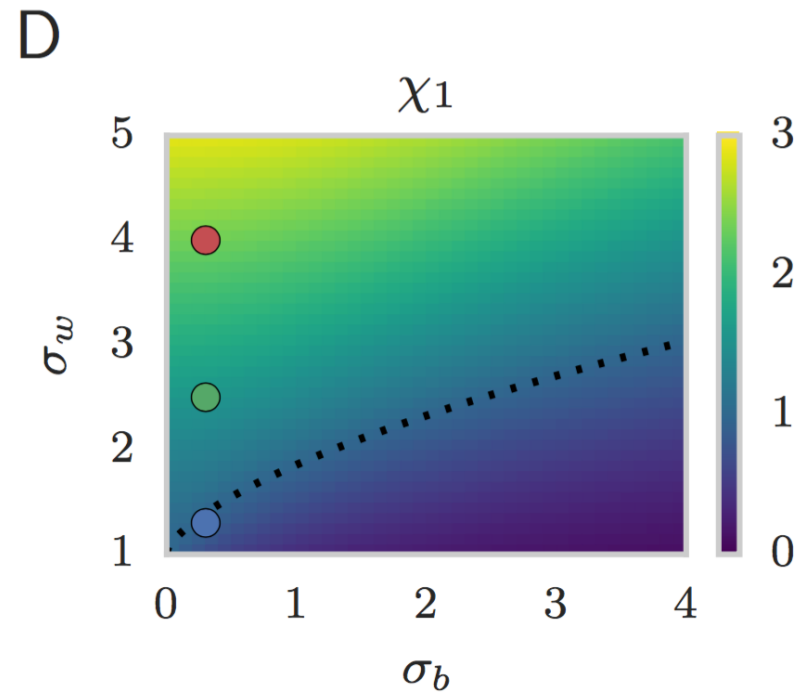
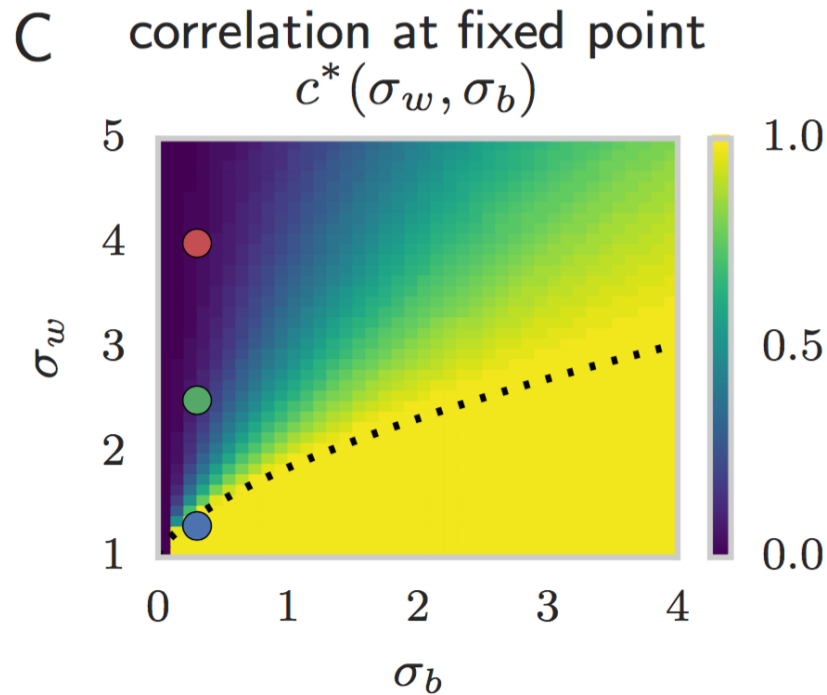
— $\sigma_w = 1.3$
 — $\sigma_w = 2.5$
 — $\sigma_w = 4.0$
 $\sigma_b = 0.3$

$$\chi_1 \equiv \left. \frac{\partial c_{12}^l}{\partial c_{12}^{l-1}} \right|_{c=1} = \sigma_w^2 \int \mathcal{D}z [\phi'(\sqrt{q^*}z)]^2$$

Interpretation: χ_1 is a multiplicative stretch factor:

- $\chi_1 < 1$: nearby points come closer together
- $\chi_1 > 1$: nearby points are driven apart

Propagation of two points through a deep network

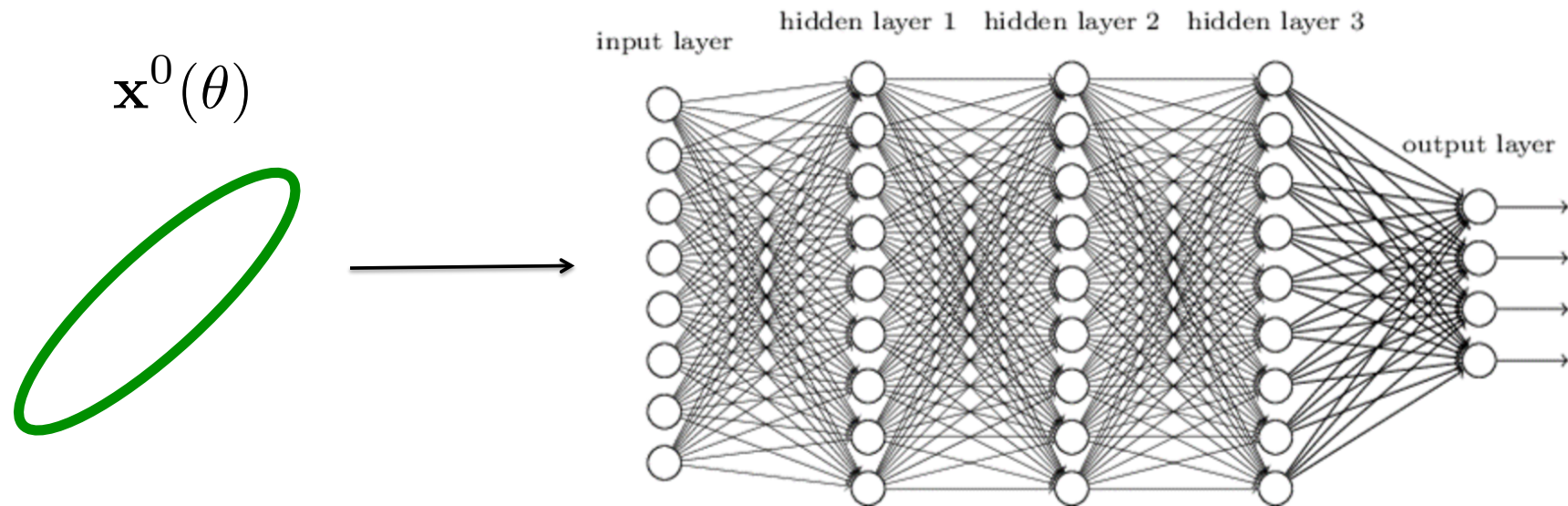


Small σ_w relative to σ_b : $\chi_1 < 1$ $c_{12}^l \rightarrow 1$

Intermediate σ_w relative to σ_b : $\chi_1 > 1$ $c_{12}^l \rightarrow c^*$

Large σ_w relative to σ_b : $\chi_1 > 1$ $c_{12}^l \rightarrow 0$

Propagation of a manifold through a deep network



The geometry of the manifold is captured by the similarity matrix -
How similar two points are in internal representation space):

$$q^l(\theta_1, \theta_2) = \frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}_i^l[\mathbf{x}^0(\theta_1)] \mathbf{h}_i^l[\mathbf{x}^0(\theta_2)]$$

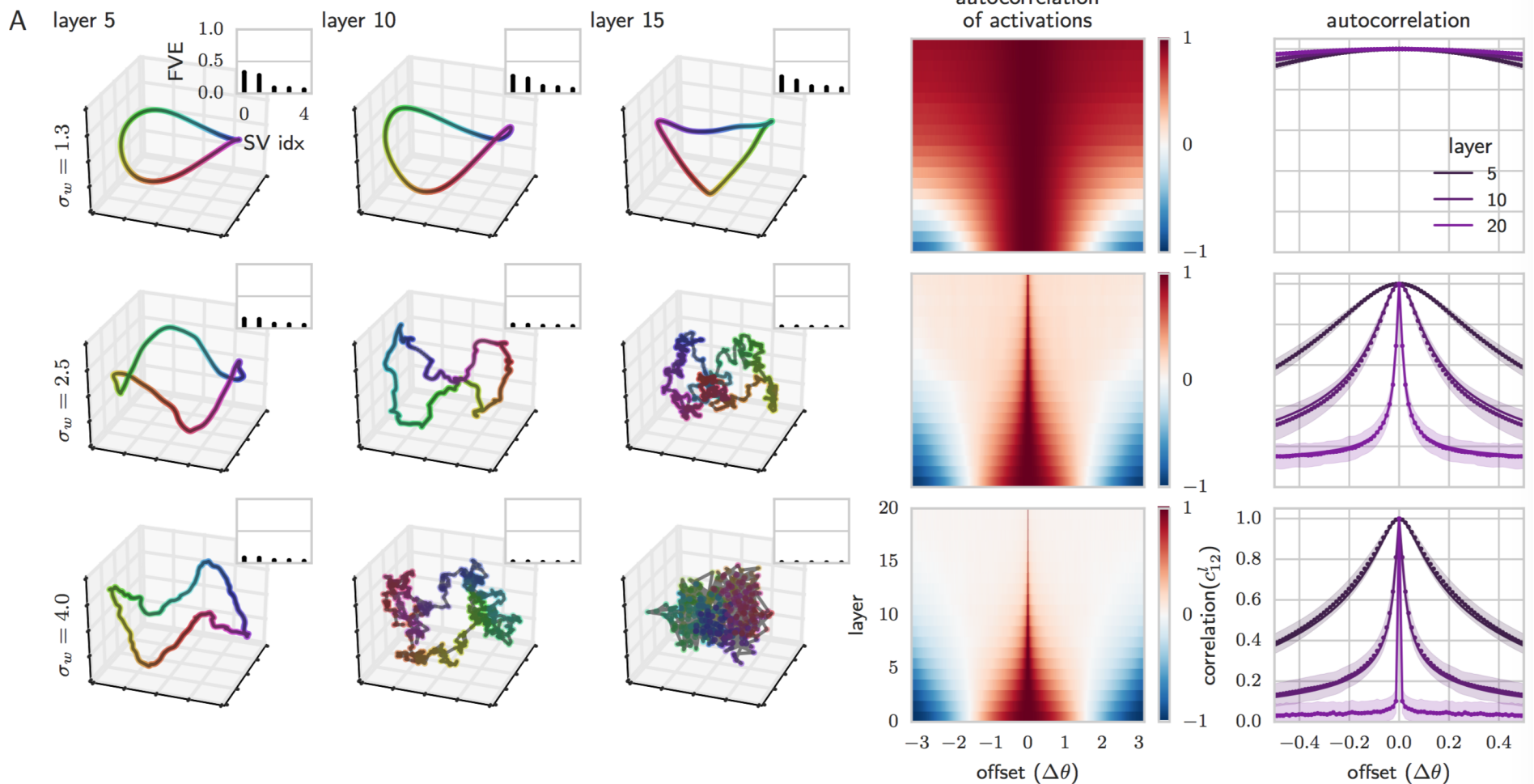
Or autocorrelation function:

$$q^l(\Delta\theta) = \int d\theta q^l(\theta, \theta + \Delta\theta)$$

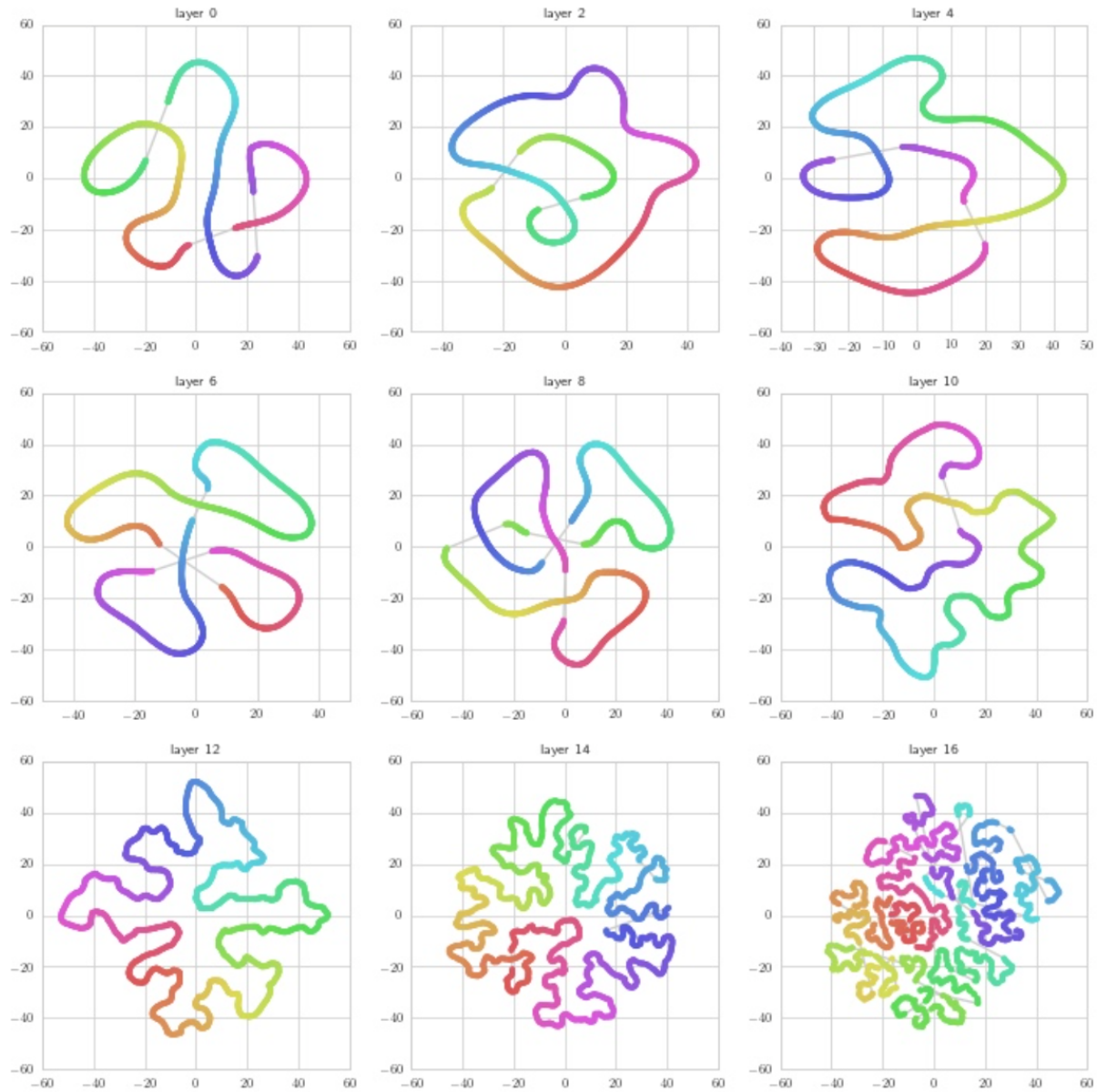
Propagation of a manifold through a deep network

$$\mathbf{h}^1(\theta) = \sqrt{N_1 q^*} [\mathbf{u}^0 \cos(\theta) + \mathbf{u}^1 \sin(\theta)]$$

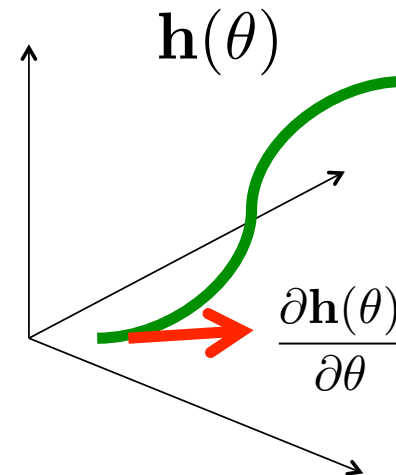
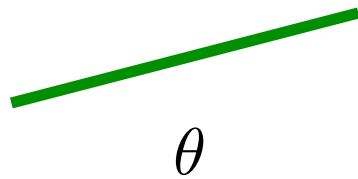
A great circle
input manifold



Propagation of a manifold through a deep network



Riemannian geometry I: Euclidean length



$$g^E(\theta) = \frac{\partial \mathbf{h}(\theta)}{\partial \theta} \cdot \frac{\partial \mathbf{h}(\theta)}{\partial \theta}$$

Metric on manifold coordinate θ induced by Euclidean metric in internal representation space \mathbf{h} .

$$d\mathcal{L}^E = \sqrt{g^E(\theta)} d\theta$$

Length element: if one moves from Θ to $\Theta + d\Theta$ along the manifold, then one moves a distance dL^E in internal representation space

Riemannian geometry II: Extrinsic Gaussian Curvature

$$\mathbf{h}(\theta)$$

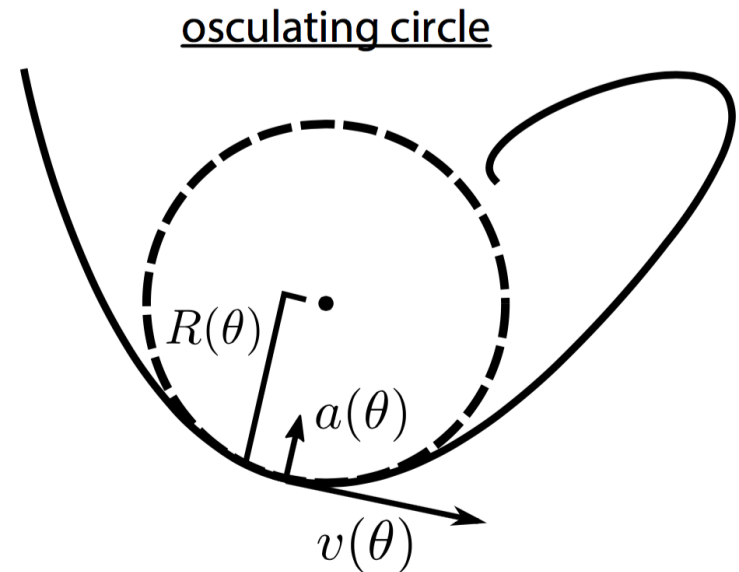
Point on the curve

$$\mathbf{v}(\theta) = \frac{\partial \mathbf{h}(\theta)}{\partial \theta}$$

Tangent or velocity vector

$$\mathbf{a}(\theta) = \frac{\partial \mathbf{v}(\theta)}{\partial \theta}$$

Acceleration vector



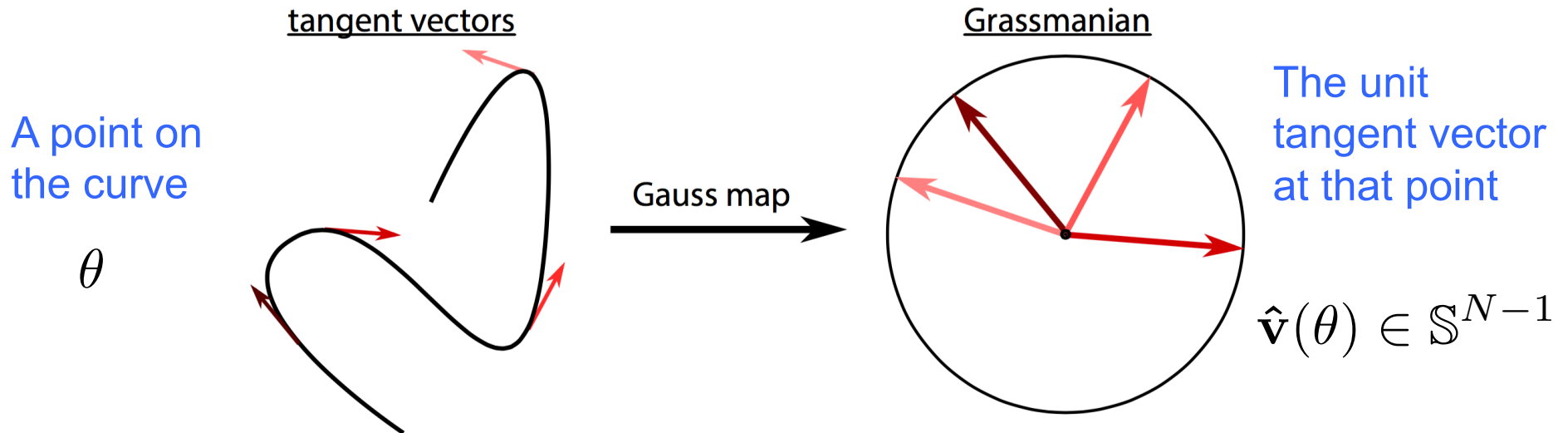
The velocity and acceleration vector span a 2 dimensional plane in N dim space.

Within this plane, there is a unique circle that touches the curve at $\mathbf{h}(\theta)$, with the same velocity and acceleration.

The Gaussian curvature $\kappa(\theta)$ is the inverse of the radius of this circle.

$$\kappa(\theta) = \sqrt{\frac{(\mathbf{v} \cdot \mathbf{v})(\mathbf{a} \cdot \mathbf{a}) - (\mathbf{v} \cdot \mathbf{a})^2}{(\mathbf{v} \cdot \mathbf{v})^3}}$$

Riemannian geometry III: The Gauss map and Grassmannian length



$$g^G(\theta) = \frac{\partial \hat{\mathbf{v}}(\theta)}{\partial \theta} \cdot \frac{\partial \hat{\mathbf{v}}(\theta)}{\partial \theta}$$

Metric on manifold coordinate θ induced by metric on the Grassmannian: how quickly unit tangent vector changes

$$d\mathcal{L}^G = \sqrt{g^G(\theta)} d\theta$$

Length element: if one moves from θ to $\theta + d\theta$ along the manifold, then one moves a distance $d\mathcal{L}^G$ Along the Grassmannian

$$g^G(\theta) = \kappa(\theta)^2 g^E(\theta)$$

Grassmannian length, Gaussian curvature and Euclidean length

An example: the great circle

$$\mathbf{h}^1(\theta) = \sqrt{Nq} [\mathbf{u}^0 \cos(\theta) + \mathbf{u}^1 \sin(\theta)]$$

A great circle
input manifold

Euclidean
length

Gaussian
Curvature

Grassmannian
Length

$$g^E(\theta) = Nq$$

$$\kappa(\theta) = 1/\sqrt{Nq}$$

$$g^G(\theta) = 1$$

$$\mathcal{L}^E = 2\pi\sqrt{Nq}$$

$$\mathcal{L}^G = 2\pi$$

Behavior under isotropic linear expansion via multiplicative stretch χ_1 :

$$\mathcal{L}^E \rightarrow \sqrt{\chi_1} \mathcal{L}^E$$

$$\kappa \rightarrow \frac{1}{\sqrt{\chi_1}} \kappa$$

$$\mathcal{L}^G \rightarrow \mathcal{L}^G$$

$\chi_1 < 1$ Contraction

Increase

Constant

$\chi_1 > 1$ Expansion

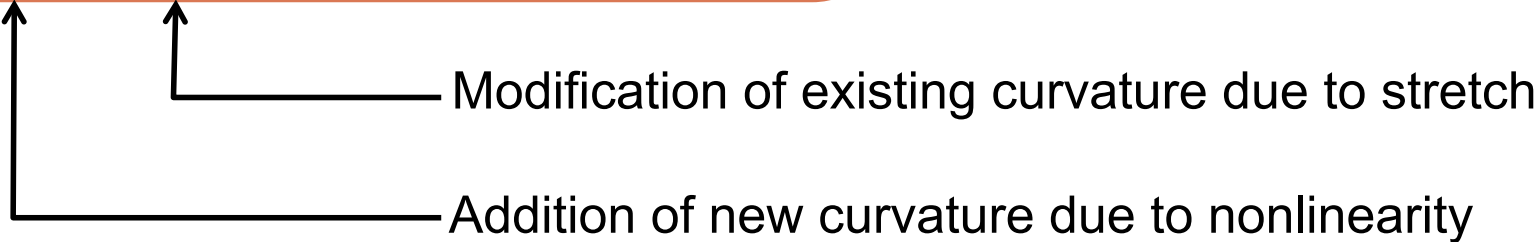
Decrease

Constant

Theory of curvature propagation in deep networks

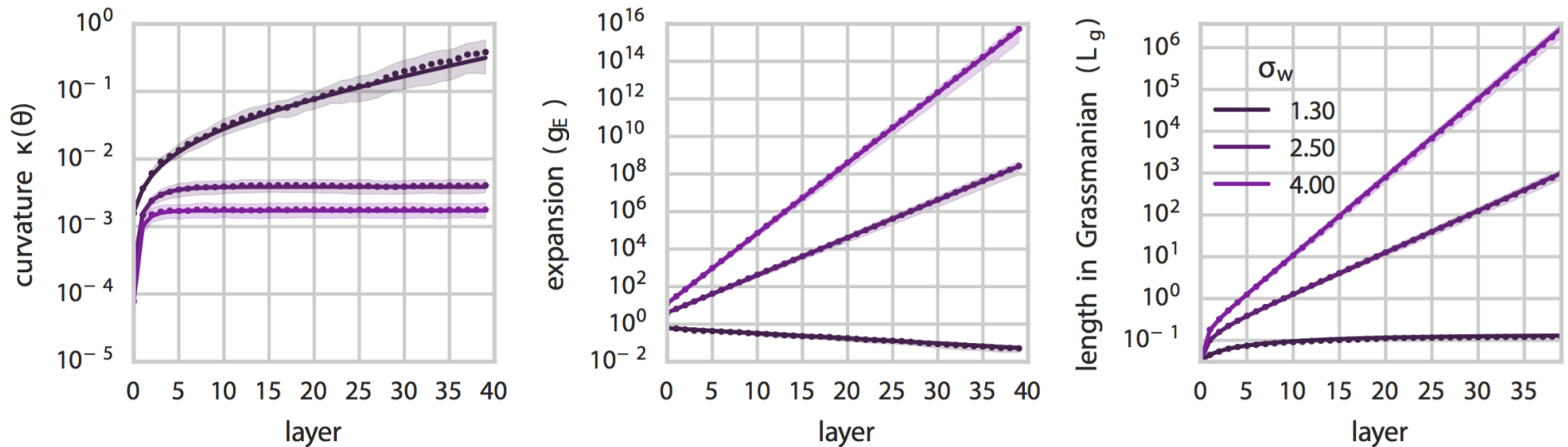
$$\begin{array}{c|c}
 \bar{g}^{E,l} = \chi_1 \bar{g}^{E,l-1} & \bar{g}^{E,1} = q^* \\
 \hline
 (\bar{\kappa}^l)^2 = 3 \frac{\chi_2}{\chi_1^2} + \frac{1}{\chi_1} (\bar{\kappa}^{l-1})^2 & (\bar{\kappa}^1)^2 = \frac{1}{q^*}
 \end{array}$$

$$\begin{array}{l}
 \chi_1 = \sigma_w^2 \int \mathcal{D}z [\phi'(\sqrt{q^*}z)]^2 \\
 \chi_2 = \sigma_w^2 \int \mathcal{D}z [\phi''(\sqrt{q^*}z)]^2
 \end{array}$$



		Local Stretch	Gaussian Curvature	Grassmannian Length
Ordered:	$\chi_1 < 1$	Contraction	Explosion	Constant
Chaotic:	$\chi_1 > 1$	Expansion	Attenuation + Addition	Exponential Growth

Curvature propagation: theory and experiment

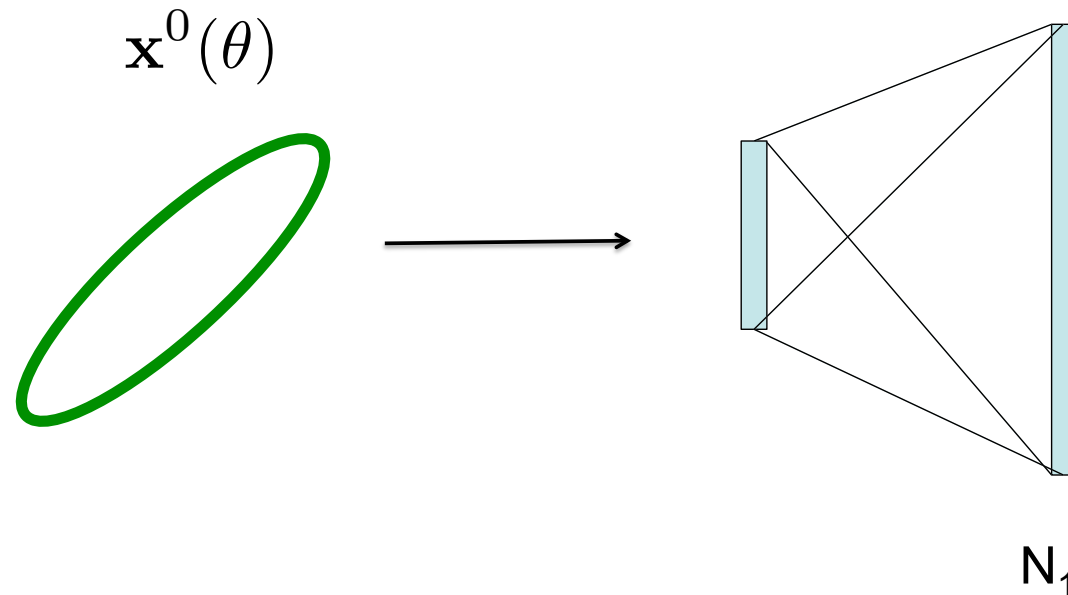


Unlike linear expansion, deep neural signal propagation can:

- 1) exponentially expand length,
- 2) without diluting Gaussian curvature,
- 3) thereby yielding exponential growth of Grassmannian length.

As a result, the circle will become space filling as it winds around at a constant rate of curvature to explore many dimensions!

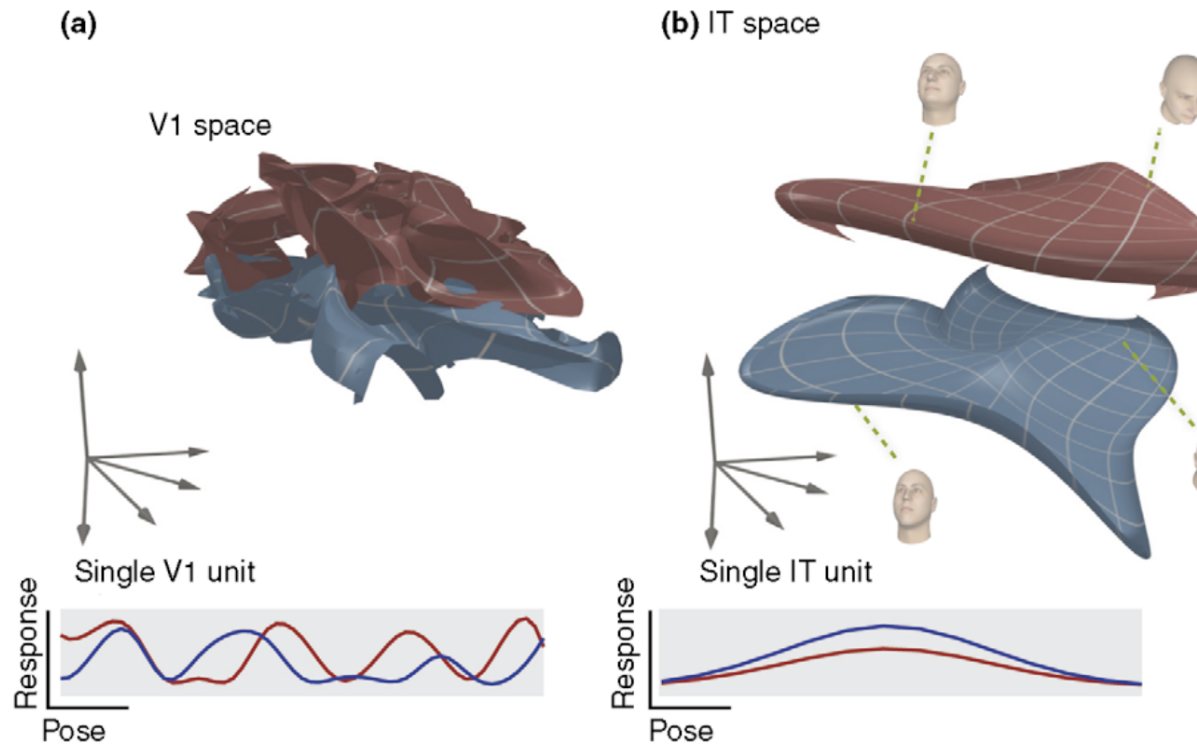
Exponential expressivity is not achievable by shallow nets



Consider a shallow network with 1 hidden layer \mathbf{x}^1 , one input layer \mathbf{x}^0 , with $\mathbf{x}^1 = \phi(\mathbf{W}^1 \mathbf{x}^0) + \mathbf{b}^1$, and a linear readout layer. How complex can the hidden representation be as a function of its width N_1 , relative to the results above for depth? We prove a general upper bound on \mathcal{L}^E (see SM):

Theorem 1. *Suppose $\phi(h)$ is monotonically non-decreasing with bounded dynamic range R , i.e. $\max_h \phi(h) - \min_h \phi(h) = R$. Further suppose that $\mathbf{x}^0(\theta)$ is a curve in input space such that no 1D projection of $\partial_\theta \mathbf{x}(\theta)$ changes sign more than s times over the range of θ . Then for any choice of \mathbf{W}^1 and \mathbf{b}^1 the Euclidean length of $\mathbf{x}^1(\theta)$, satisfies $\mathcal{L}^E \leq N_1(1 + s)R$.*

Boundary disentangling: theory



How can we mathematically formalize the notion of disentangling in deep networks?

How do we use this mathematical formalization to quantitatively assess the disentangling power of deep versus shallow networks?

Boundary disentangling: theory

$$y = \text{sgn}(\boldsymbol{\beta} \cdot \mathbf{x}^D - \beta_0)$$

$$(\boldsymbol{\beta} \cdot \mathbf{x}^D - \beta_0) = 0$$

A linear classifier in the top layer

Implements a hyperplane decision boundary in final layer

$$G(\mathbf{x}^0) = (\boldsymbol{\beta} \cdot \mathbf{x}^D(\mathbf{x}^0) - \beta_0) = 0$$

Yielding a curved co-dimension 1 decision boundary in the input layer

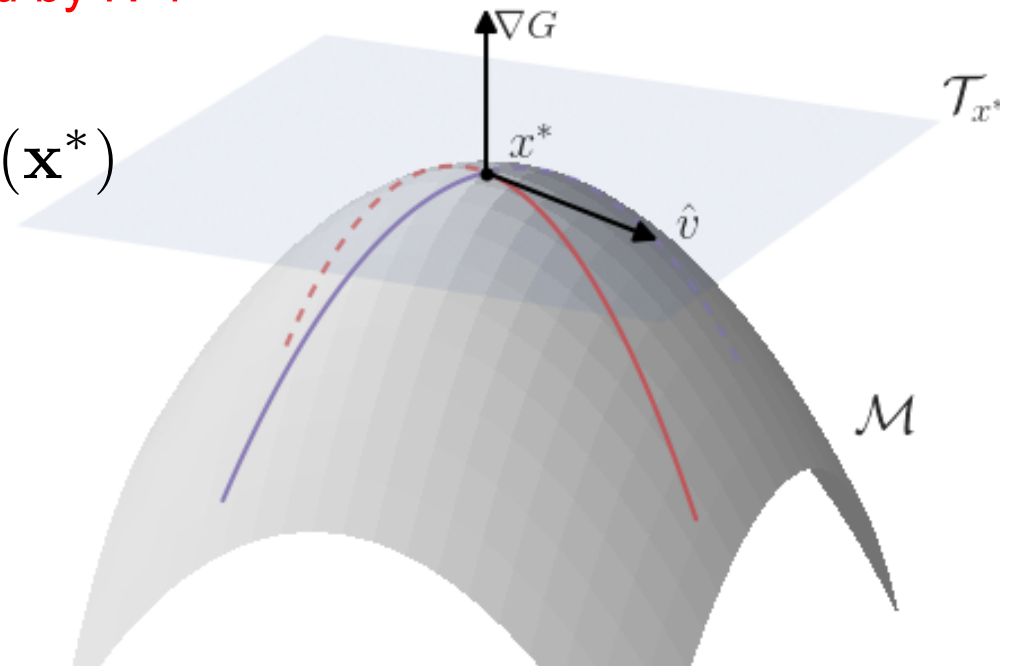
Its curvature at a point is characterized by N-1 principal curvatures:

$$\kappa_1(\mathbf{x}^*) \geq \kappa_2(\mathbf{x}^*) \geq \dots \geq \kappa_{N-1}(\mathbf{x}^*)$$

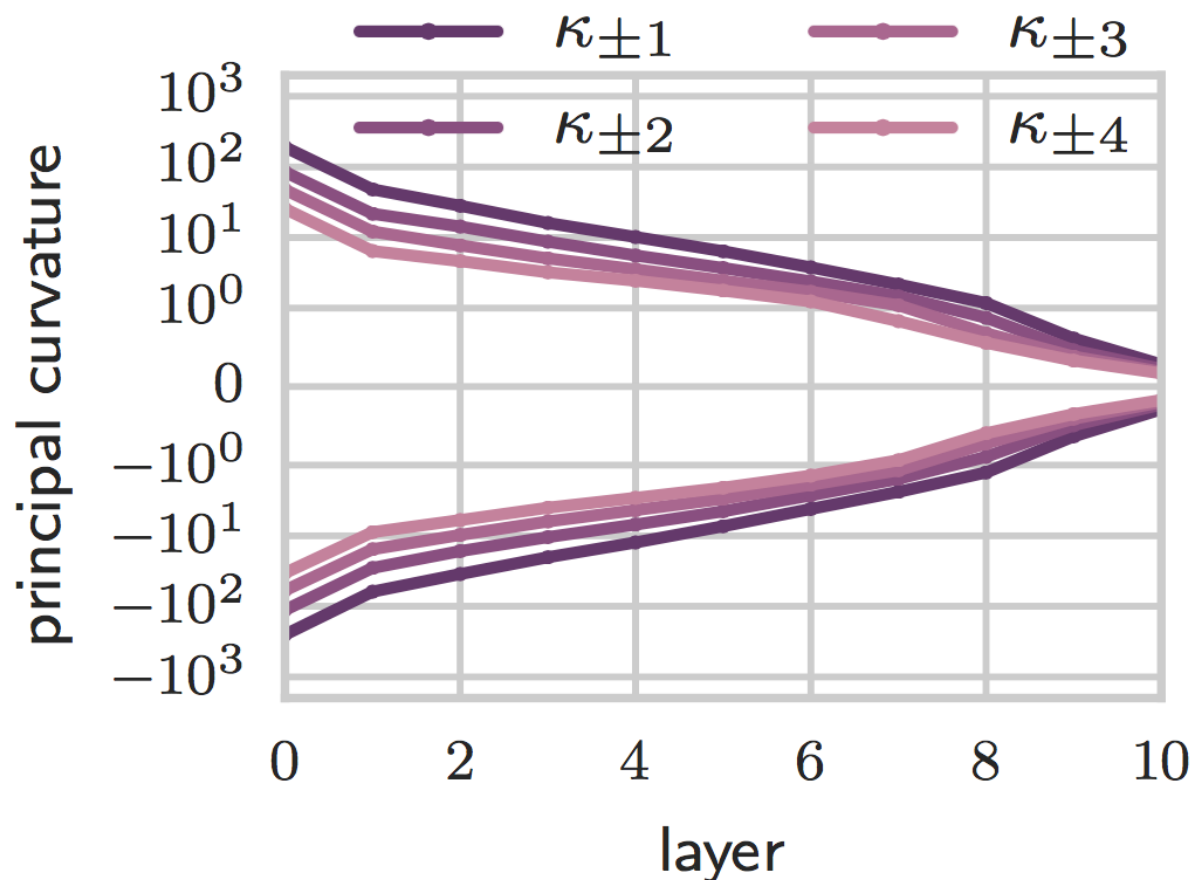
They are the eigenvalues of:

$$\mathcal{H} = \|\vec{\nabla} G\|_2^{-1} \mathbf{P} \frac{\partial^2 G}{\partial \mathbf{x} \partial \mathbf{x}^T} \mathbf{P}$$

$$\mathbf{P} = \mathbf{I} - \widehat{\nabla G} \widehat{\nabla G}^T$$



Boundary disentangling: experiment



The principal curvatures of decision boundaries in the chaotic regime grow **exponentially** with depth!

Thus exponentially curved manifolds in input space can be flattened to hyperplanes even by deep random networks!

Summary

We have combined Riemannian geometry with dynamical mean field theory to study the emergent deterministic properties of signal propagation in deep nonlinear nets.

We derived analytic recursion relations for Euclidean length, correlations, curvature, and Grassmannian length as simple input manifolds propagate forward through the network.

We obtain an excellent quantitative match between theory and simulations.

Our results reveal the existence of a transient chaotic phase in which the network expands input manifolds without straightening them out, leading to “space filling” curves that explore many dimensions while turning at a constant rate. The number of turns grows exponentially with depth.

Such exponential growth does not happen with width in a shallow net.

Chaotic deep random networks can also take exponentially curved $N-1$ Dimensional decision boundaries in the input and flatten them into Hyperplane decision boundaries in the final layer: exponential disentangling!

Some of the theoretical puzzles of deep learning

Trainability: if a good network solution exists with small training error, how do we find it? And what makes a learning problem difficult?

A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks ICLR 2014.

A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, CogSci 2013.

Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, under review, NIPS 2016.

Generalizability: what principles do deep networks use to place probability / make decisions in regions of input space with little data?

M. Advani and S. Ganguli, Statistical Mechanics of Optimal Convex Inference in High Dimensions, Physical Review X, 2016.

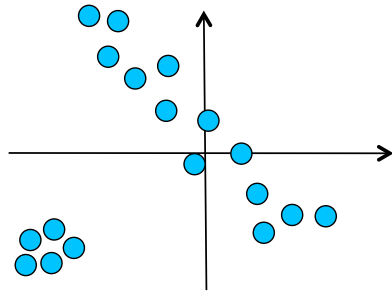
Expressiveness, Memorization, Stability, and Flat versus sharp minima.

Statistical mechanics of high dimensional data analysis

N = dimensionality of data M = number of data points

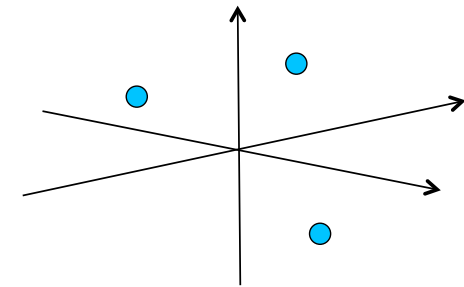
$\alpha = N / M$

Classical Statistics



$N \sim O(1)$
 $M \rightarrow \infty$
 $\alpha \rightarrow 0$

Modern Statistics



$N \rightarrow \infty$
 $M \rightarrow \infty$
 $\alpha \sim O(1)$

Machine Learning and Data Analysis

Learn statistical parameters by maximizing log likelihood of data given parameters.

Statistical Physics of Quenched Disorder

Energy = $-\log \text{Prob}(\text{data} | \text{parameters})$
Data = quenched disorder
Parameters = thermal degrees of freedom

Statistical mechanics of compressed sensing, S. Ganguli and H. Sompolinsky, PRL 2010.

Short-term memory in neuronal networks through dynamical compressed sensing, NIPS 2010.

Compressed sensing, sparsity and dimensionality in neuronal information processing and data analysis, S. Ganguli and H. Sompolinsky, Annual Reviews of Neuroscience, 2012

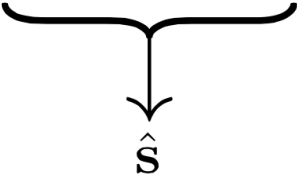
Statistical mechanics of optimal convex inference in high dimensions, M. Advani and S. Ganguli, Physical Review X, 2016.

An equivalence between high dimensional Bayes optimal inference and M-estimation, NIPS 2016.

Random projections of random manifolds, S. Lahiri, P. Gao, S. Ganguli, <http://arxiv.org/abs/1607.04331>.

Optimal inference in high dimensions

$$s^0 \longrightarrow \mathbf{x}_\mu \longrightarrow y_\mu = \mathbf{x}_\mu \cdot \mathbf{s}^0 + \epsilon_\mu$$



Generative model and measurements

P dim signal $s^0 \sim P_s$

N measurements with noise $\epsilon \sim P_\epsilon$

$\alpha = N/P =$ measurement density

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s}} \sum_{\mu} \rho(\mathbf{y}_\mu - \mathbf{x}_\mu \cdot \mathbf{s}) + \sum_j \sigma(s_j)$$

Estimation algorithm

$\rho =$ loss function

$\sigma =$ regularizer

$q_s =$ L_2 estimation error

$$\frac{1}{P} \sum_j (\hat{s}_j - s_j^0)^2 = q_s(\alpha, \rho, \sigma, P_\epsilon, P_s)$$

Least squares: $\rho(\epsilon) = \epsilon^2$

$\sigma(s) = 0$

Maximum likelihood: $\rho(\epsilon) = -\log P_\epsilon(\epsilon)$

$\sigma(s) = 0$

Ridge regression: $\rho(\epsilon) = \epsilon^2$

$\sigma(s) = s^2$

LASSO: $\rho(\epsilon) = \epsilon^2$

$\sigma(s) = \lambda_1 |s|$

Elastic Net: $\rho(\epsilon) = \epsilon^2$

$\sigma(s) = \lambda_1 |s| + \lambda_2 s^2$

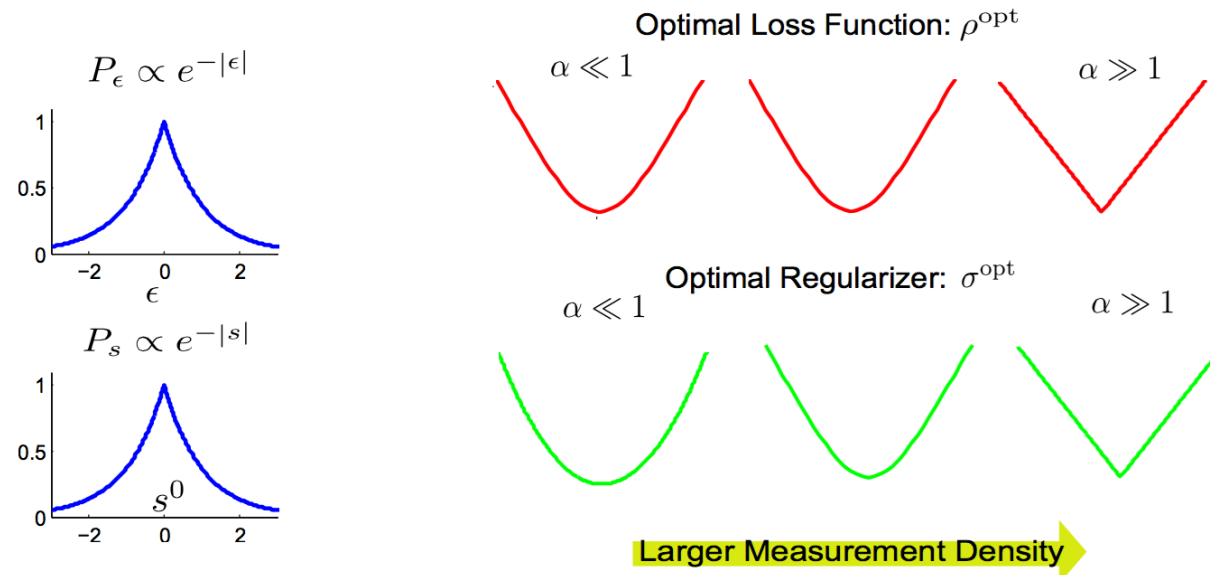
MAP: $\rho(\epsilon) = -\log P_\epsilon(\epsilon)$

$\sigma(s) = -\log P_s(s)$

Example algorithms

Optimal inference in high dimensions

Question: For a given signal distribution P_s , noise distribution P_ϵ , and measurement density α , what is the best loss function ρ and regularizer σ ?



For log-concave signal and noise: the optimal loss and regularizer are nonlinearly smoothed versions of MAP where the smoothing **increases** as the measurement density **decreases**.

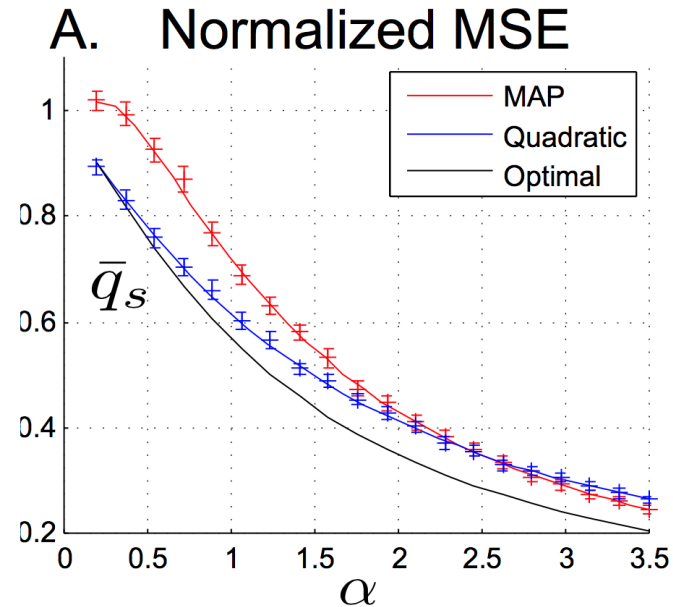
MAP is optimal at **high** measurement density.

Ridge regression is optimal at **low** measurement density **independent** of signal and noise!

No inference algorithm can out-perform our optimal algorithm!

Optimal inference in high dimensions

Question: For a given signal distribution P_s , noise distribution P_ε , and measurement density α , what is the best loss function ρ and regularizer σ ?



M. Advani and S. Ganguli, An equivalence between high dimensional Bayes optimal inference and M-estimation, NIPS 2016.

M. Advani and S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions, Physical Review X, 6, 031034, 2016.

Also prior work by the groups of Montanari and El-Karoui

For log-concave signal and noise: the optimal loss and regularizer are nonlinearly smoothed versions of MAP where the smoothing **increases** as the measurement density **decreases**.

MAP is optimal at **high** measurement density.

Ridge regression is optimal at **low** measurement density **independent** of signal and noise!

No inference algorithm can out-perform our optimal algorithm!

More generally: upper bounds on generalization error

Complexity based upper bounds:

$\mathcal{R}_n =$ Rademacher Complexity

$$\epsilon_{\text{gen}} \leq \epsilon_{\text{train}} + \mathcal{R}_n$$

How well you memorize a data set with random labels of size n .

Perfect memorization = 1

For linear classes, as n becomes larger than dimension, $\mathcal{R}_n \rightarrow O(1/n^{1/2})$

Stability based upper bounds:

$$\epsilon_{\text{gen}} \leq \epsilon_{\text{train}} + \epsilon(\text{w/o example } i) - \epsilon(\text{w/example } i)$$

If your learned function is robust to changes in the dataset, then you will not over fit!

Recent observations on generalization in deep nets

Complexity based upper bounds: $\mathcal{R}_n =$ Rademacher Complexity

$$\epsilon_{\text{gen}} \leq \epsilon_{\text{train}} + \mathcal{R}_n$$

How well you memorize a data set with random labels of size n .

Perfect memorization: $\mathcal{R}_n = 1$

For linear classes, as n becomes larger than dimension, $\mathcal{R}_n \rightarrow O(1/n^{1/2})$

Zhang et. al. Understanding deep learning requires rethinking generalization.

Arpit et. al. A closer look at memorization in deep Networks

Stability based upper bounds:

$$\epsilon_{\text{gen}} \leq \epsilon_{\text{train}} + \epsilon(\text{w/o example } i) - \epsilon(\text{w/example } i)$$

If your learned function is robust to changes in the dataset, then you will not over fit!

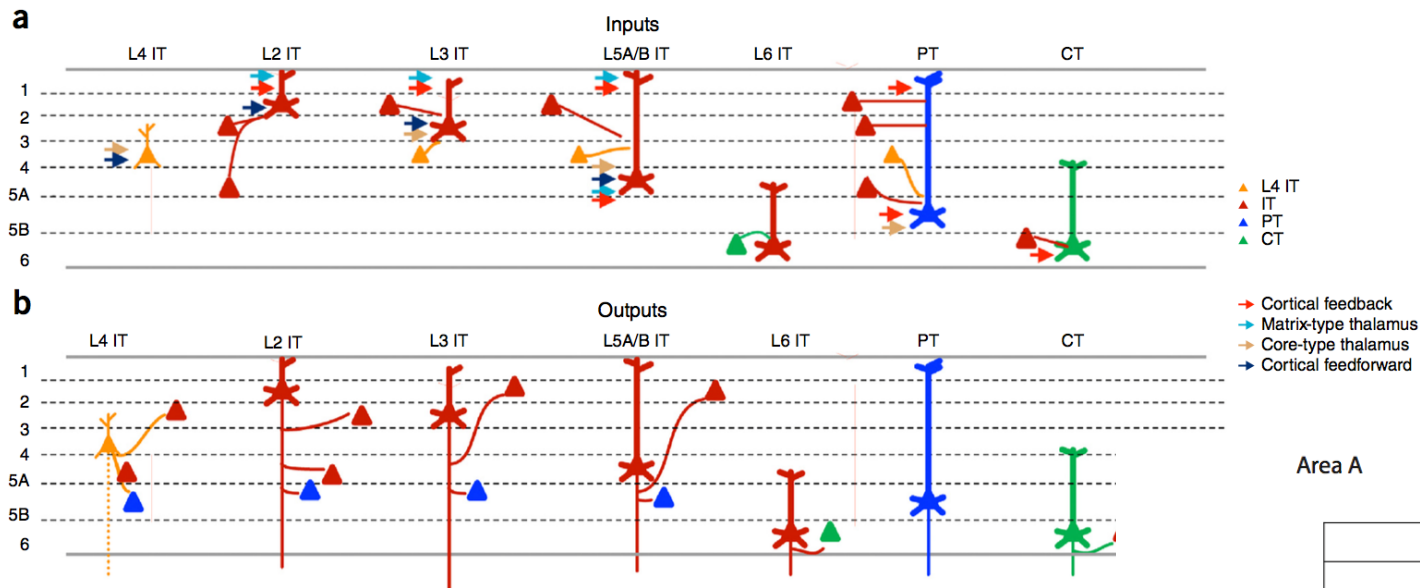
Keskar et. al. On large batch training for deep learning: generalization gap and sharp minima.

Dinh et. al. Sharp minima can generalize for deep nets.

Talk Outline

- **Applying deep learning to the brain:**
 - Recurrent neural networks for context dependent decision making
 - Feed-forward networks for modeling the ventral visual stream
 - State of the art models of retinal function
- **Theory of deep learning:**
 - Optimization
 - Expressivity
 - Generalization
- **Inspiration from neuroscience back to deep learning:**
 - Canonical cortical microcircuits
 - Nested loop architectures
 - Avoiding catastrophic forgetting through synaptic complexity
 - Learning asymmetric recurrent generative models

There are more things in heaven and earth...

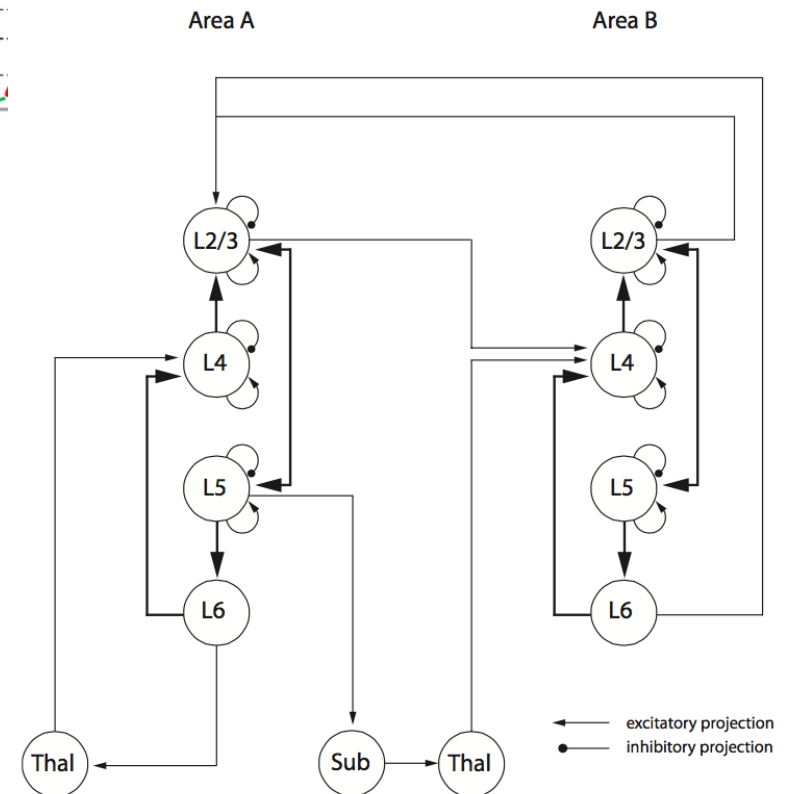


Douglas and Martin, Canonical circuits of the neocortex, *Ann. Rev. Neurosci* 2004.

Da Costa and Martin, Whose cortical column would that be? *Front. In Neuroanatomy*, 2010.

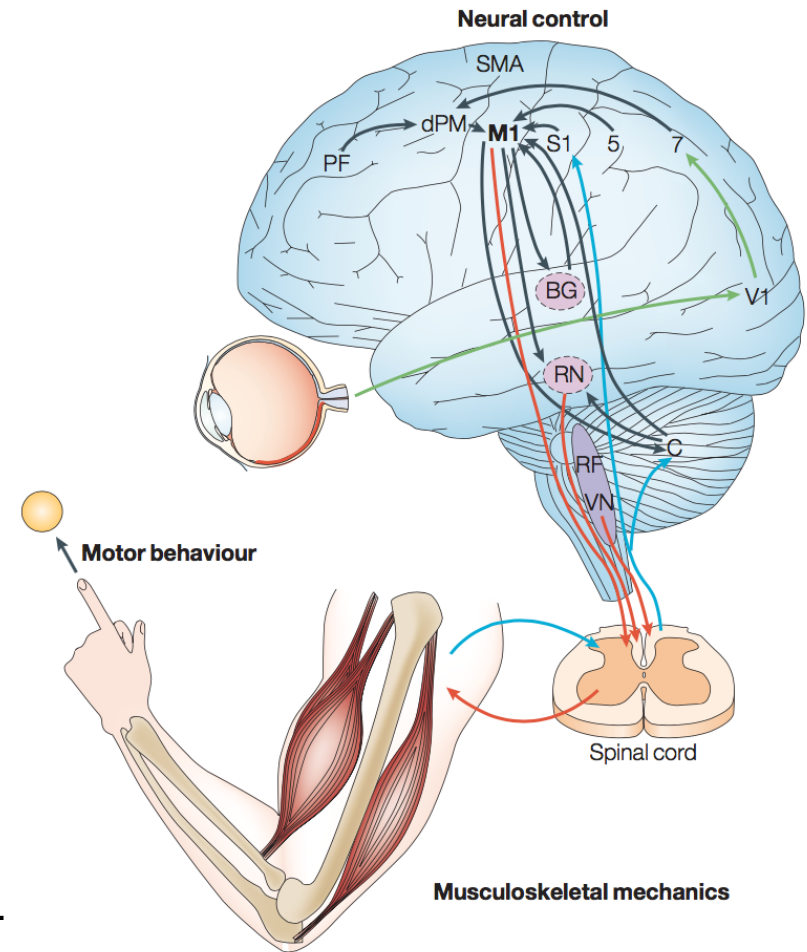
Harris and Shephard, The neocortical circuit: Themes and variation, *Nat. Neuro* 2015

Shephard, Synaptic organization of the brain, 5th ed., 2009



There are more things in heaven and earth...

Exploration of nested loop architectures



Scott, Optimal Feedback Control and the Neural Basis of Volitional Control, Nature Neurosci. 2004.

Todorov, Optimality principles in sensorimotor control, Nature Neurosci 2004.

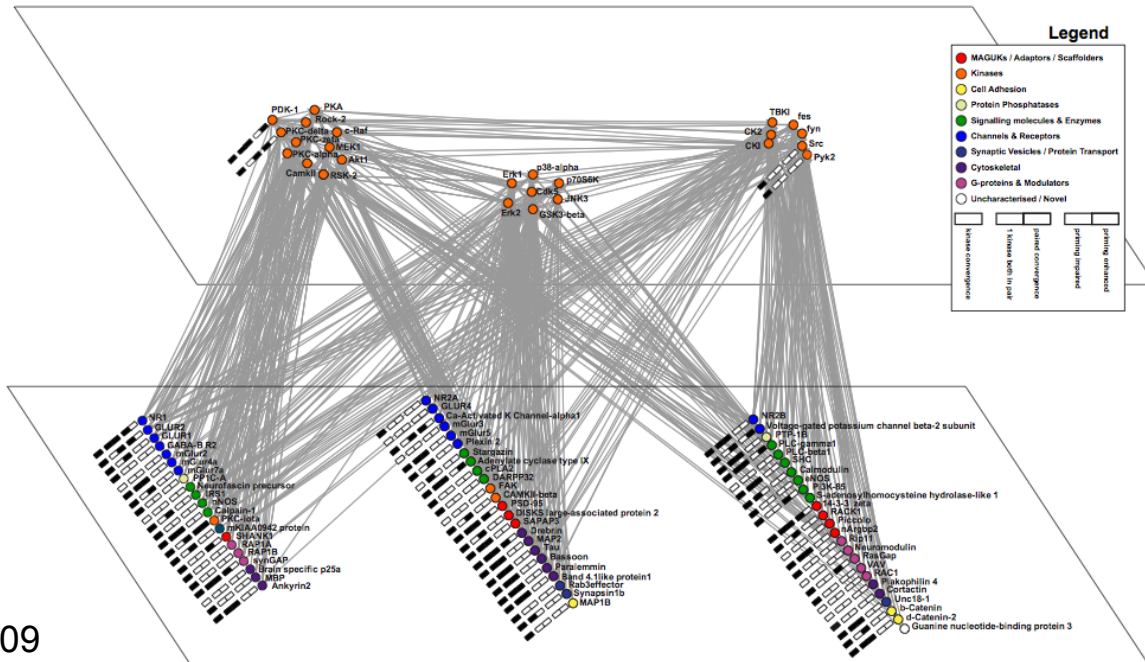
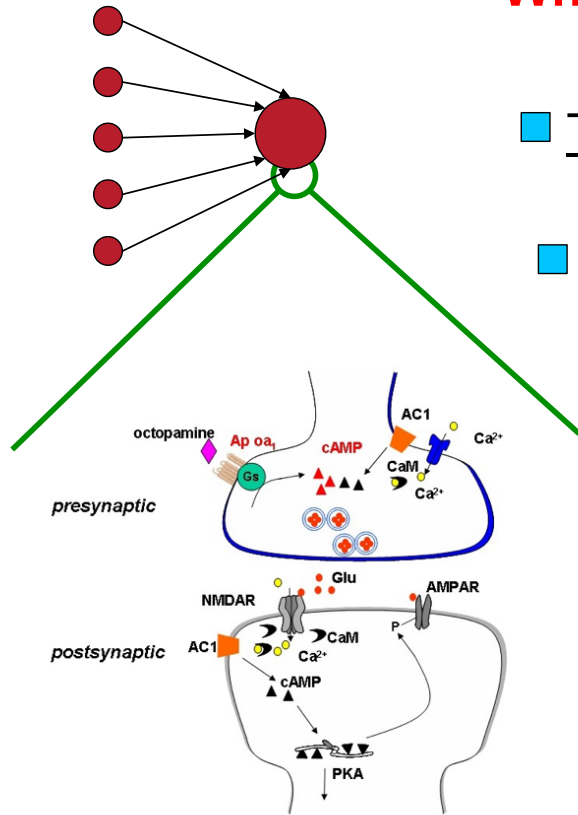
Scott, The computational and neural basis of voluntary motor control and planning, Trends in Cog. Sci 2012.

There are more things in heaven and earth...

What is a synapse from neuron j to neuron i?

■ Theorist: W_{ij} or $J_{ij} \sim$ size of postsynaptic potential

■ Experimentalist: AMPA, NMDA, CAMKII, MAPK, CREB, MHC-I, second messengers, membrane protein regulation, intracellular trafficking, new protein synthesis



Coba et. al.
Science Signalling 2009

The functional contribution of synaptic complexity to learning and memory

Shatz Lab

Han-Mi Lee

Raymond Lab

Barbara Nguyen-Vu
Grace Zhao
Aparna Suvrathan

Ganguli Lab

Subhaneil Lahiri



Funding:

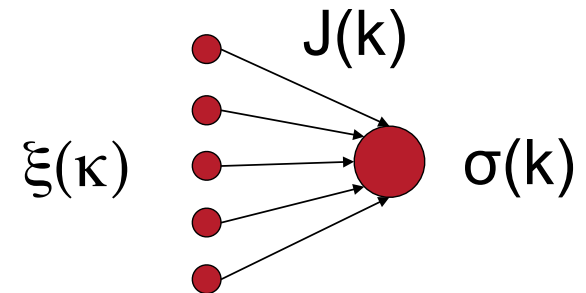
Bio-X Neuroventures
Burroughs Wellcome
Genentech Foundation
James S. McDonnell Foundation
McKnight Foundation
National Science Foundation

Office of Naval Research
Simons Foundation
Sloan Foundation
Swartz Foundation
Terman Award

Memory capacity with scalar analog synapses

Consider the number of associations a neuron with N afferent synapses can store.

$$\sigma(k) = \text{sgn} (J \cdot \xi(k) - \theta)$$



An online learning rule to store the desired association:

$$J(k+1) = e^{-1/\tau} J(k) + \sigma(k) \xi(k)$$

- i.e. 1) Allows analog weights to decay slightly (forget the past inputs)
2) Add in the new association to the weight (learn a new input).

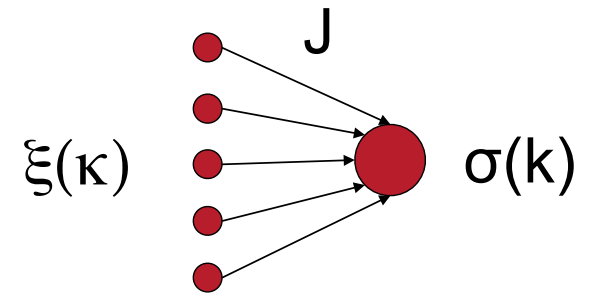
Memory capacity: How far back into the past can synapses reliably recall previously stored associations?

Answer: If τ is $O(N)$ then the past $O(N)$ associations can be recalled.

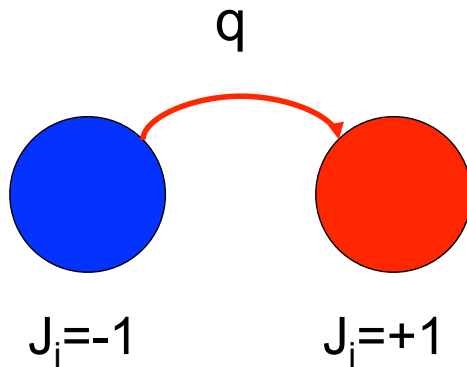
Problem: This solution relies on individual synapses to reliably maintain $O(N)$ distinguishable analog states.

Memory capacity with binary synapses

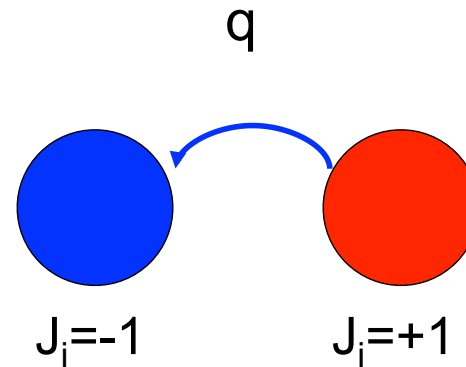
What about real synapses which can take only a finite number of distinguishable values for their strength?



For binary synapses each synapse $J_i = +1$ or -1 . So you can no longer add an association to synaptic weights without running into boundaries.

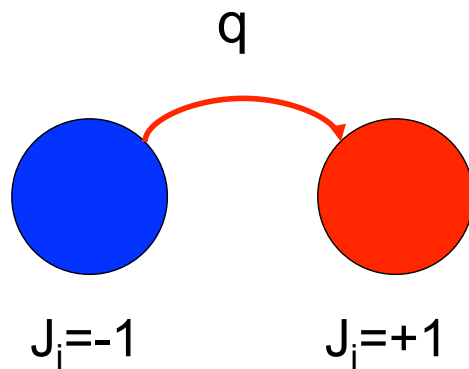


Potentiation

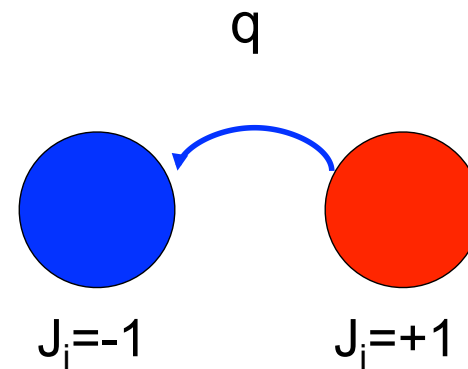


Depression

Memory capacity with binary synapses



Potentiation



Depression

q = prob a synapse changes strength under appropriate conditions
 N = number of synapses

Memory Capacity

$$q = O(1)$$

$$q = O(N^{-1/2})$$

$$\log N$$

$$N^{1/2}$$

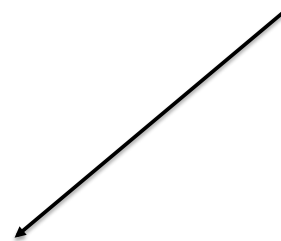
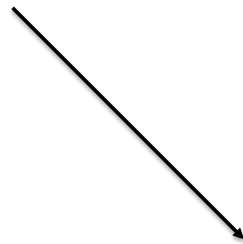
Quickly learn, quickly forget

Sluggish to learn, slow to forget

Synaptic complexity: from scalars to dynamical systems

Experiment

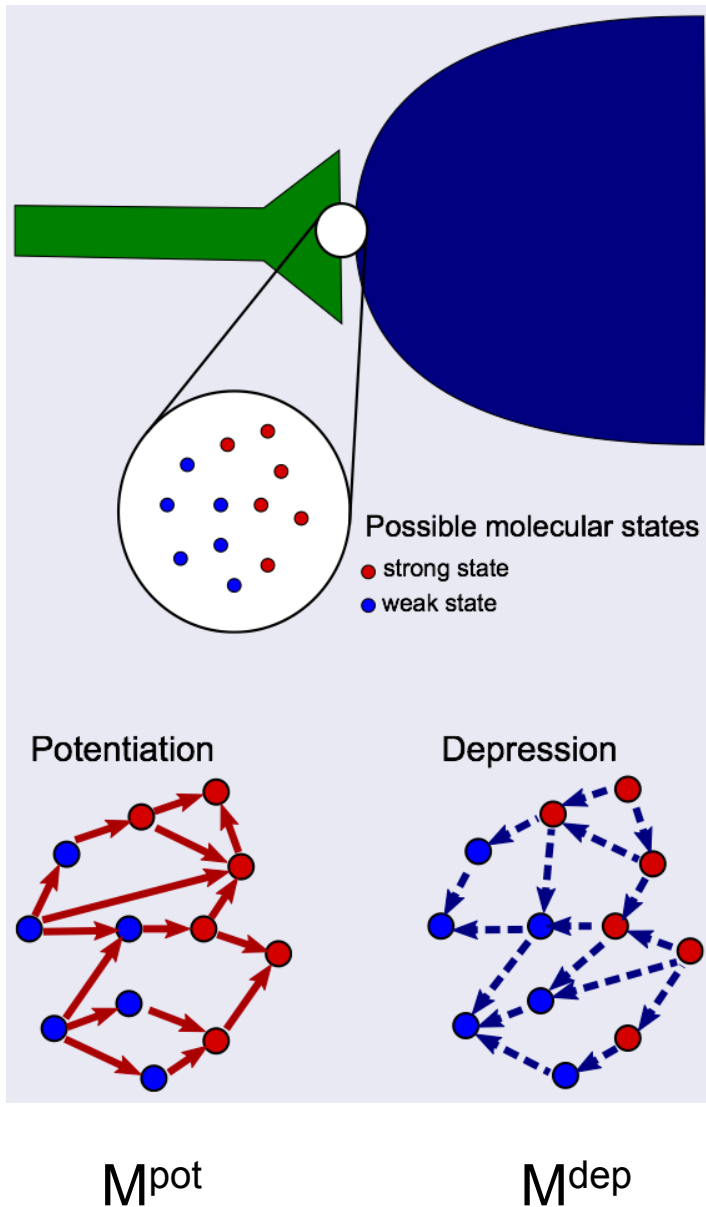
Theory



We must expand our theoretical conception of a synapse from that of a simple scalar value to an entire (stochastic) dynamical system in its own right.

This yields a large universe of synaptic models to explore and understand.

Framework for synaptic dynamical systems

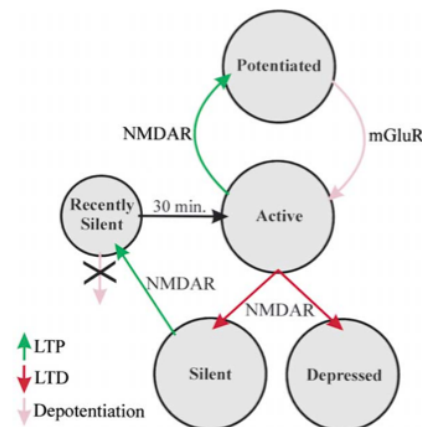


Theoretical approach:

A synapse is an arbitrary stochastic dynamical system with M internal states.

Some internal states correspond to a **strong** synapse, others a **weak** synapse.

A candidate **potentiation** (**depression**) event induces an arbitrary stochastic transition between states.



Montgomery
and Madison
Neuron
2002

Ideal observer measure of memory capacity: SNR

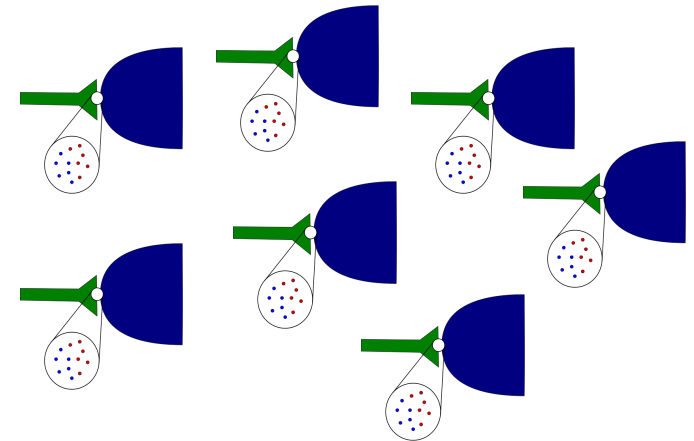
A continuous stream of memories are stored (at poisson rate r) in a population of N synapses with M internal states.

The memory stored at time $t=0$ demands that some synapses potentiate, while others depress, yielding an ideal synaptic weight vector w_{ideal} .

The storage of future memories after $t=0$ changes the weight vector to $w(t)$.

An upper bound on the quality of memory retrieval of any memory readout using neural activity is given by the SNR curve:

$$\text{SNR}(t) = \frac{\langle \vec{w}_{\text{ideal}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\text{ideal}} \cdot \vec{w}(\infty) \rangle}{\sqrt{\text{Var}(\vec{w}_{\text{ideal}} \cdot \vec{w}(\infty))}}$$



Each choice of

N , M , M^{pot} and M^{dep}

yields a different memory curve.

Two example synaptic molecular networks

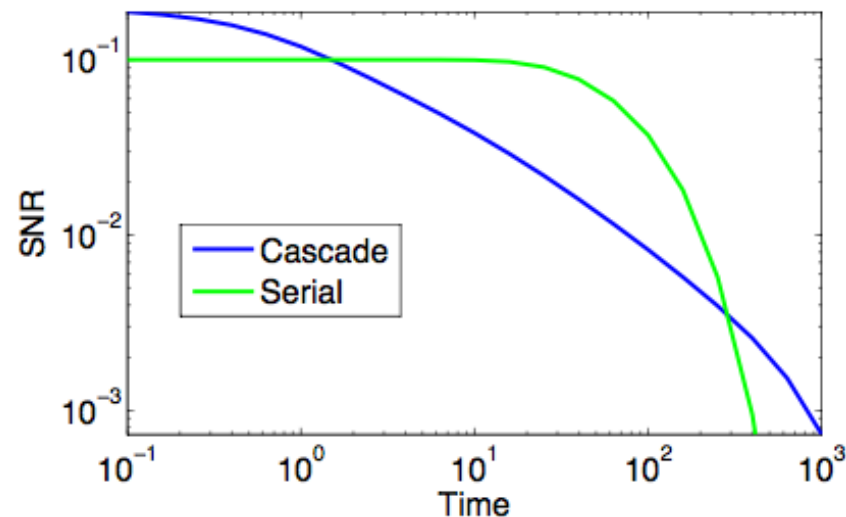
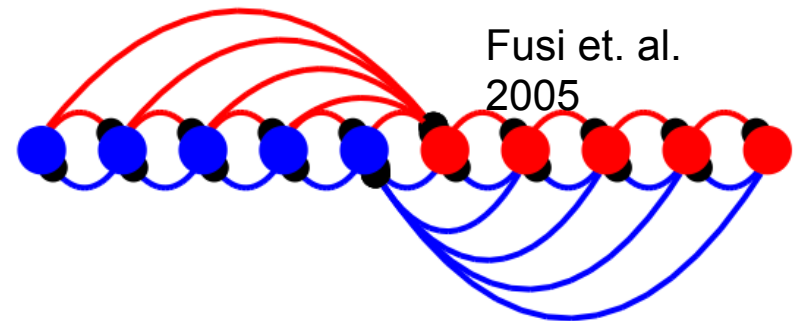
Serial Model

Leibold and Kempter
2008



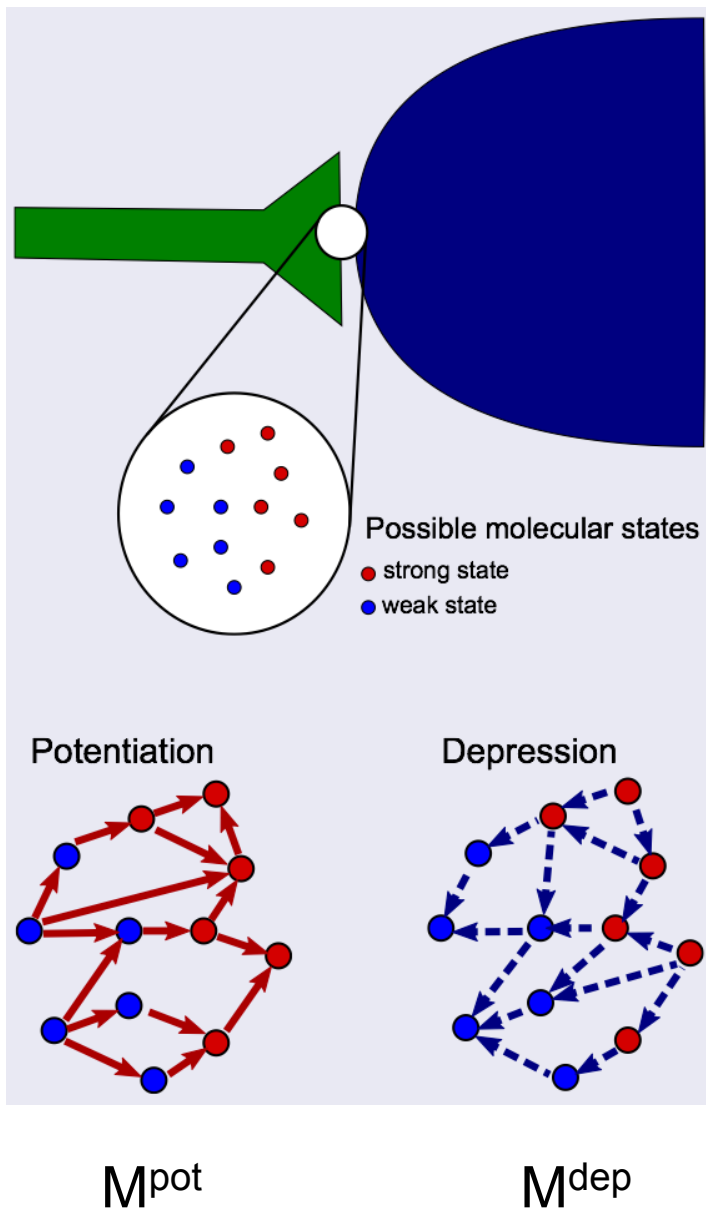
Cascade Model

Fusi et. al.
2005



To elucidate the functional contribution of molecular complexity to memory, we want to not simply understand individual models, but understand the space of all possible models within this family.

Towards a general theory of synaptic complexity



How does the structure of a synaptic dynamical system (M^{pot} and M^{dep}) determine its function, or memory curve $\text{SNR}(t)$?

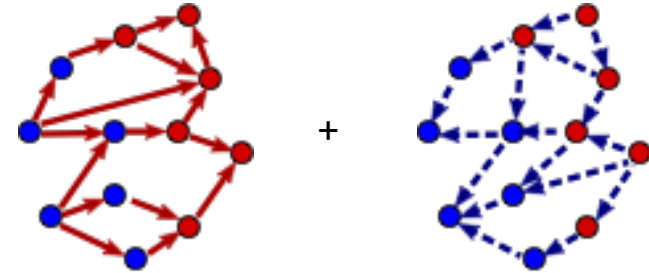
What are the fundamental limits of achievable memory over all possible choices of synaptic dynamical systems?

What is the structural organization of synaptic dynamical systems that achieve these limits?

What theoretical principles can control combinatorial explosion in the number of possible models as M increases?

Imposing a theoretical order on synaptic dynamics

As the synaptic population undergoes continuous modification, the internal state stochastically wanders around according to a forgetting process:



$$M^{\text{forget}} = f^{\text{pot}} * M^{\text{pot}} + f^{\text{dep}} * M^{\text{pot}}$$

This forgetting process has:

An equilibrium probability distribution of state occupancy: p_i^{∞}

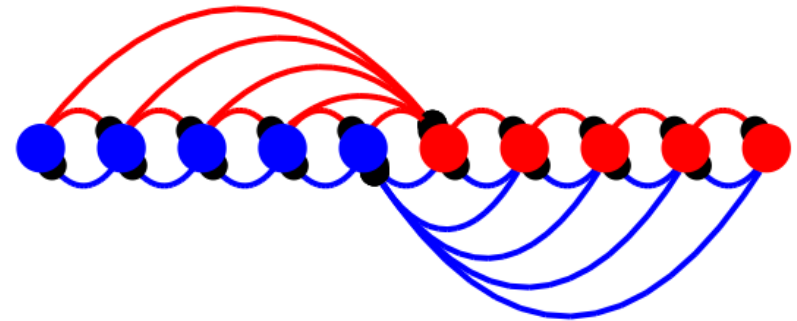
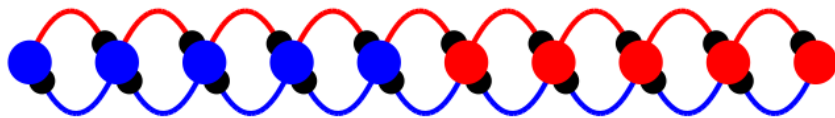
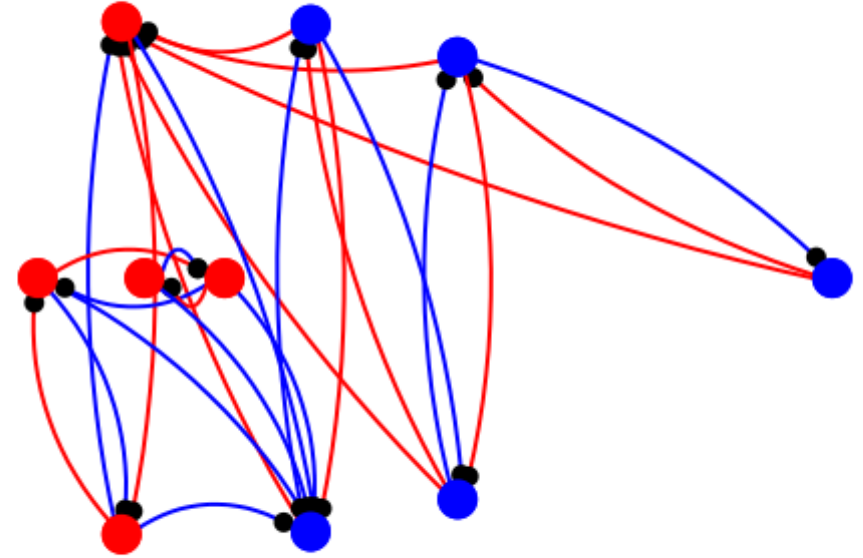
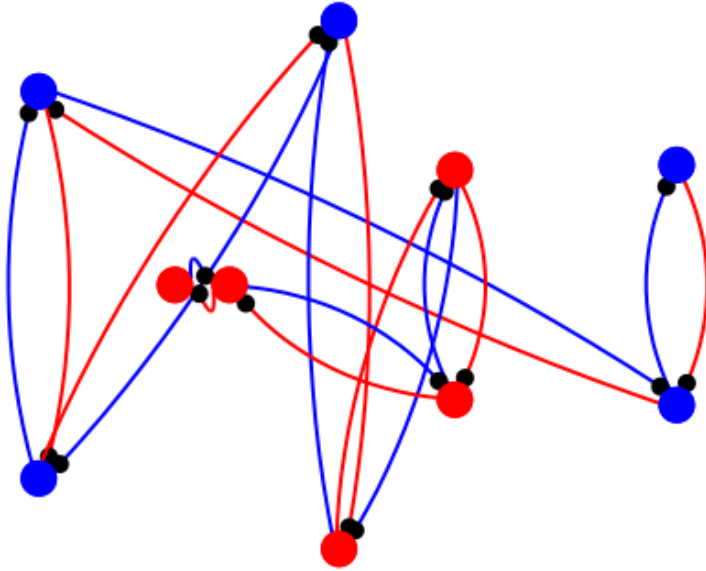
And a mean first passage time matrix from state i to j : T_{ij}

$$\eta_i^{\text{pot}} \equiv \sum_{j \in \text{pot}} T_{ij} p_j^{\infty}$$

Starting from state i , the average time it takes to get to the potentiated states, weighted by their equilibrium probability.

Order states from left to right in order of decreasing η_i^{pot}

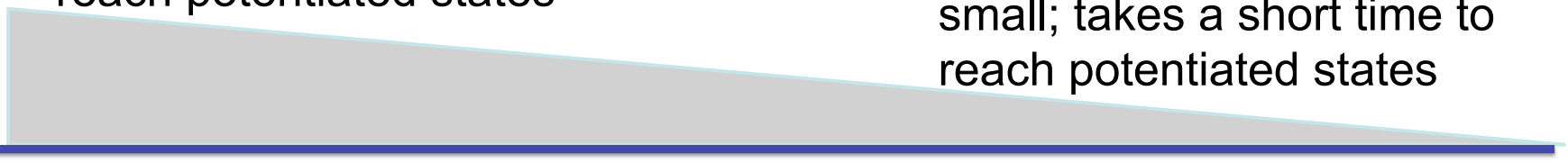
Topological ordering from first passage times



large; takes a long time to reach potentiated states

small; takes a short time to reach potentiated states

η_i^{pot}



Optimal synapses have a simple structure in this order

Consider optimizing the area under the memory curve:

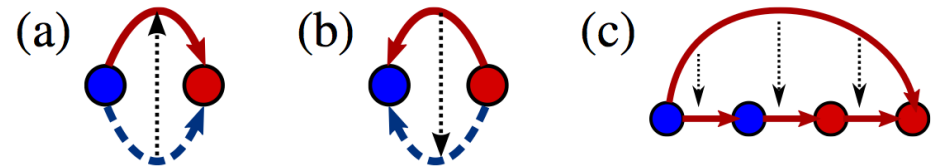
When states are placed in this order,

(a) M^{pot} should only go from left to right

(b) M^{dep} should only go from right to left

(c) We can remove shortcuts in both M^{pot} and M^{dep} while

- (1) preserving the order
- (2) preserving the equilibrium distribution
- (3) increasing the area



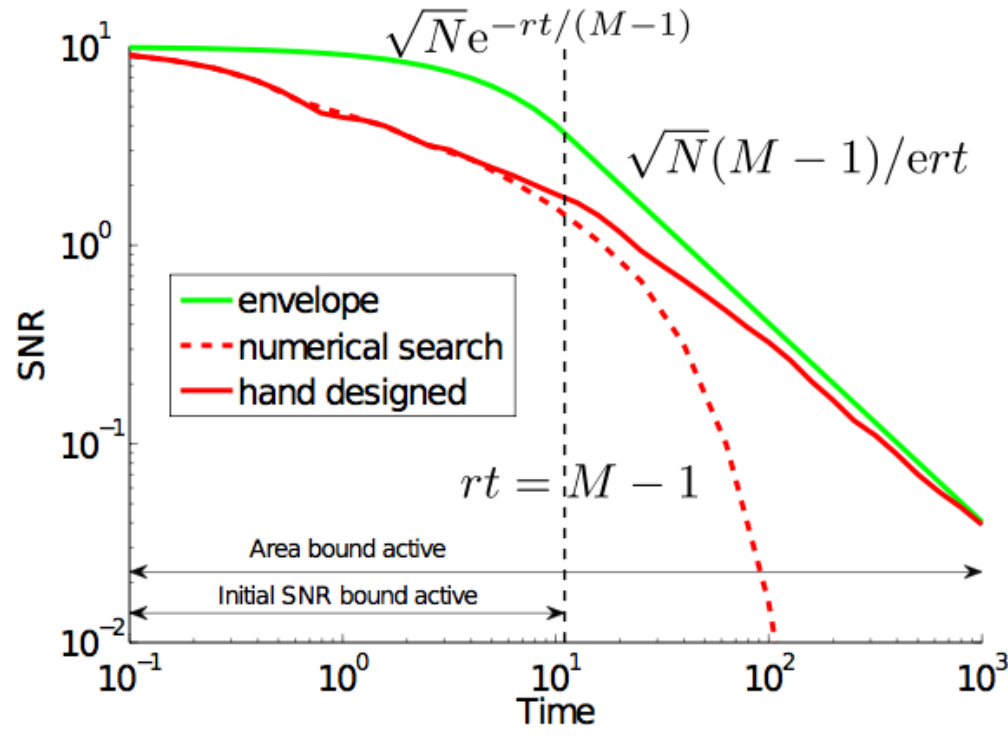
⇒ The area under the memory curve of any synaptic dynamical system is bounded by that of a chain with the same equilibrium distribution.

Also, we show that the area of a chain cannot exceed $O(N^{1/2} M)$ for any choice of transition rates along the chain.

⇒ The area under the memory curve of any synaptic dynamical system can never exceed $O(N^{1/2} M)$.

A frontier beyond whose bourn no curve can cross

Area bound implies a maximal achievable memory at any finite time given N synapses with M internal states:



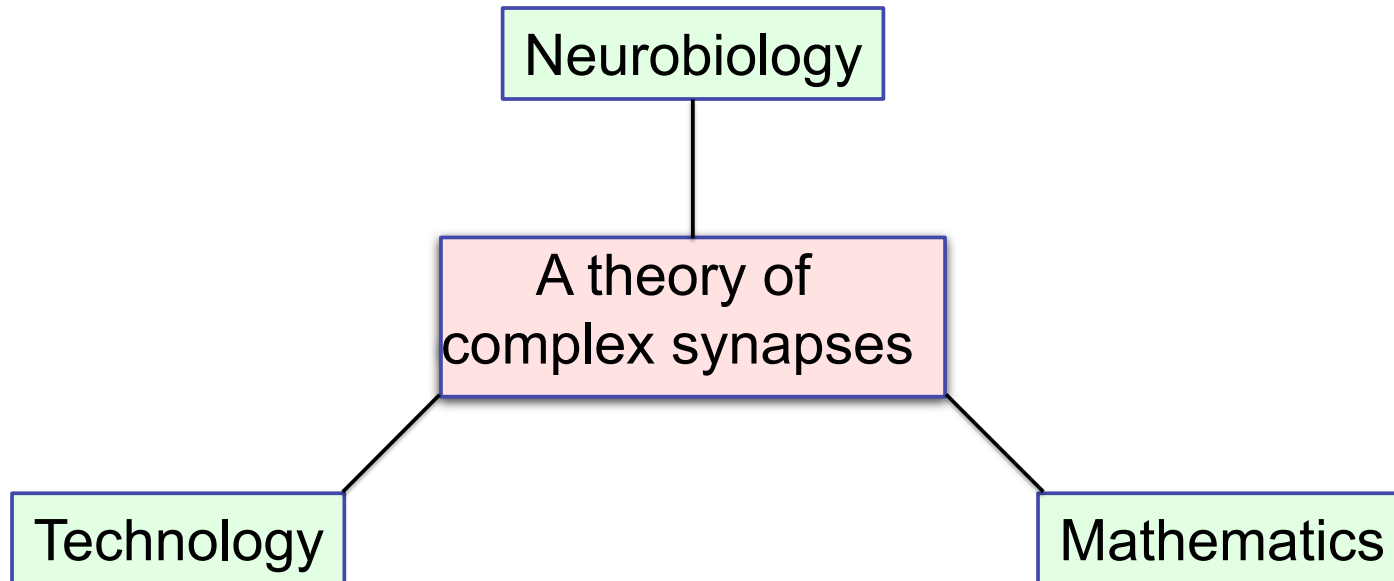
Chains with different transition rates come close to the frontier at late times.

Various measures of memory (area, frontier, lifetime) grow linearly with the number of internal states M , but grow only as the square root of the number of synapses N .

The dividends of understanding synaptic complexity

(Under review: cerebellar learning with complex synapses)

A framework for interpreting
molecular neurobiology data



The next generation of
artificial neural networks?

(Spatiotemporal credit assignment)
(Learning as message passing)

New theorems about
perturbations
to stochastic processes.

(Tighter bounds)

A potential route to cognitive enhancement?

Enhance synaptic plasticity

Enhance learning

Tang et. al. Nature 1999
Malleret et. al. Cell 2001
Guan et. al. Nature 2009

Impair Learning

Migaud et. al. Nature 1998
Hayashi et. al. Neuron 2004
Koekkoek et. al. Neuron 2005

Shatz Lab

Knockout MHC-I in
cerebellum

Enhanced LTD

Raymond Lab

Measure WT and KO
VOR learning

Observe **both** enhanced
and impaired learning

Ganguli Lab

Theoretical framework
to elucidate principles
of plasticity sufficient
to explain learning
patterns

Continual learning through synaptic intelligence

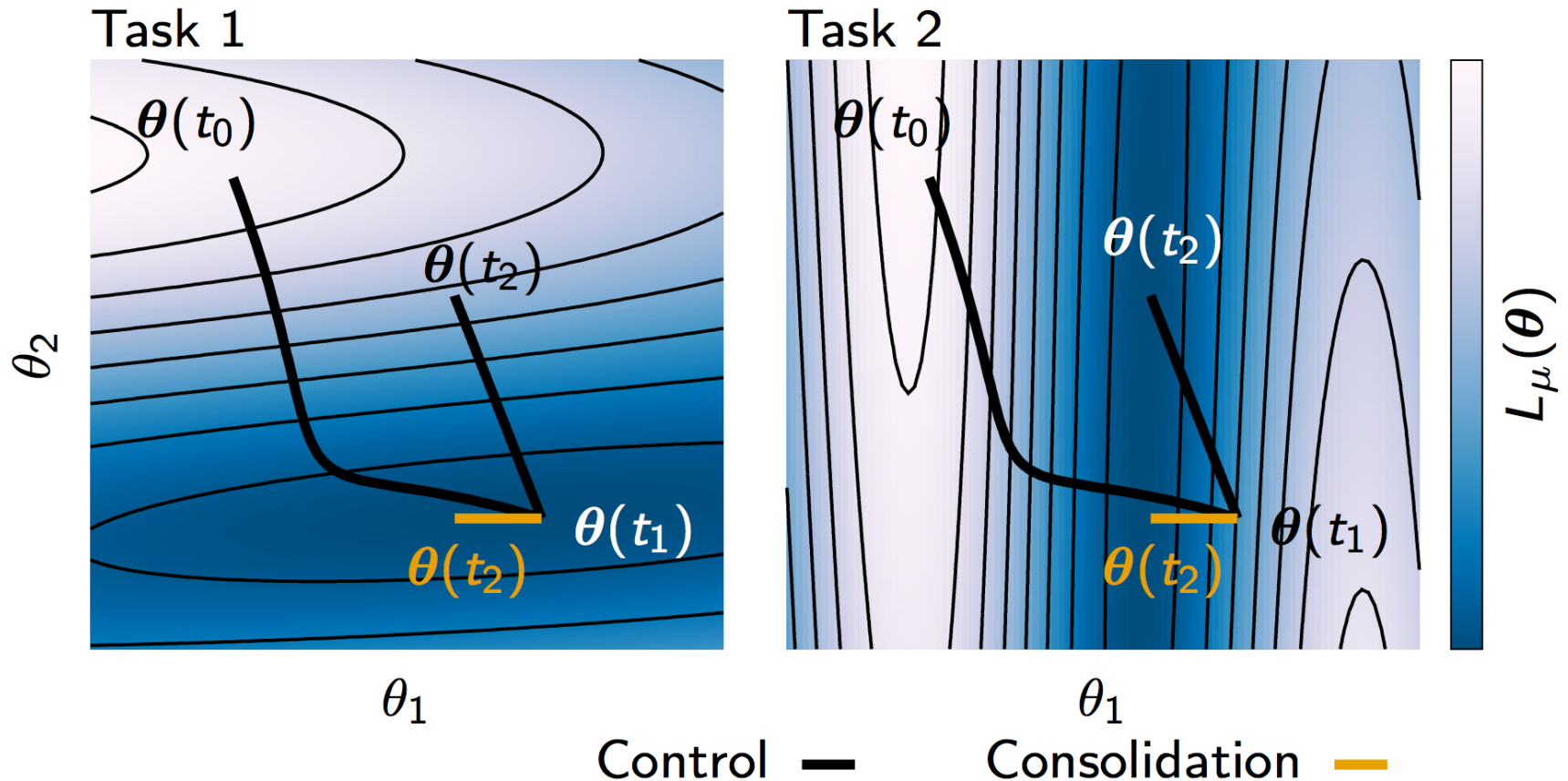


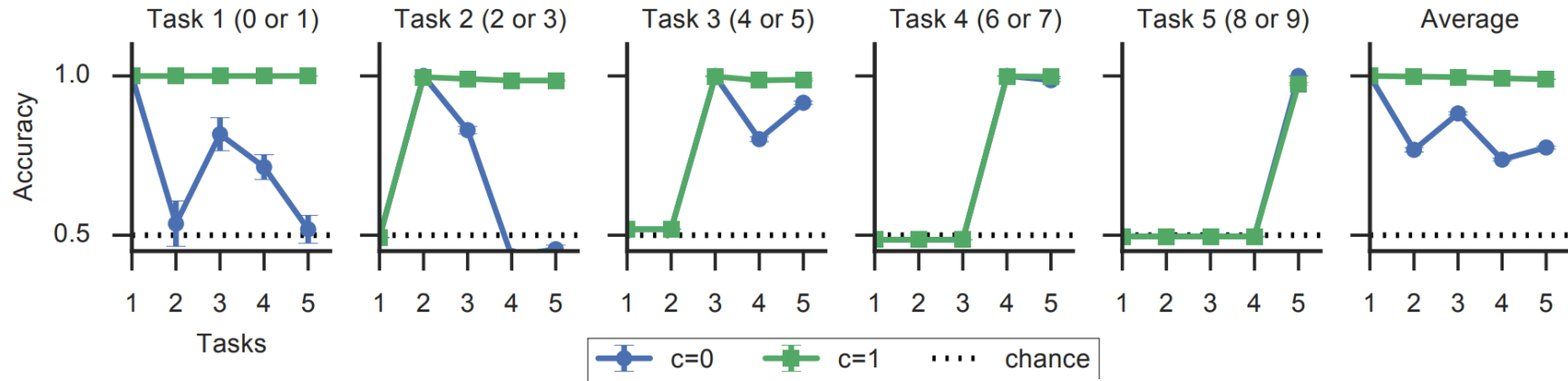
Illustration of catastrophic forgetting: solving task 2 impairs learning on solving task 1.

Idea: each synapse computes its “importance” in solving previous tasks. In future tasks unimportant synapses are allowed to change.

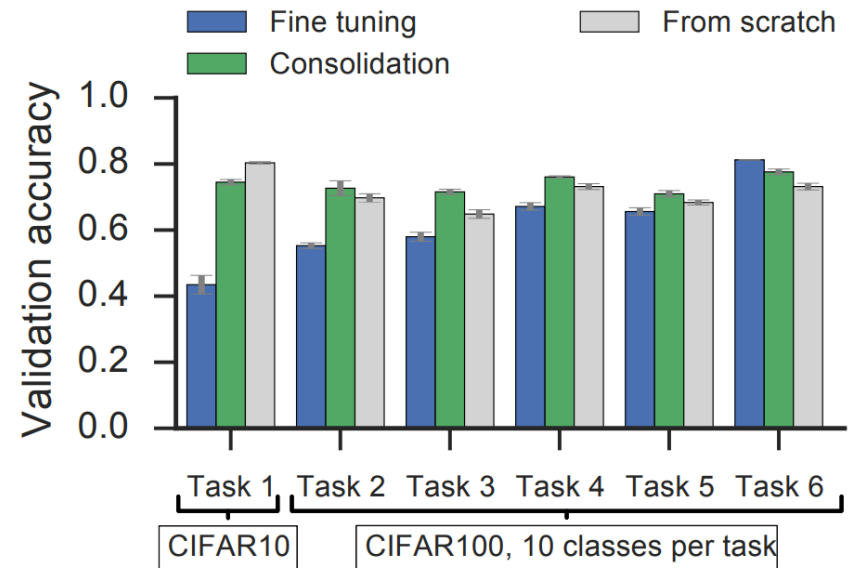
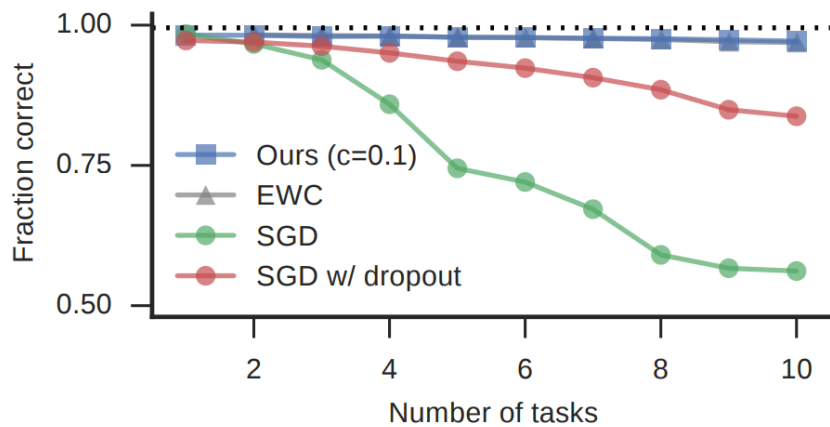
Friedemann Zenke, Ben Poole, Surya Ganguli Continual Learning Through Synaptic Intelligence, ICML 2017.

Continual learning through synaptic intelligence

Split MNIST



Permuted MNIST



Friedemann Zenke, Ben Poole, Surya Ganguli Continual Learning Through Synaptic Intelligence, ICML 2017.

Summary

Trainability: if a good network solution exists with small training error, how do we find it? And what makes a learning problem difficult?

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

Generalizability: what principles do deep networks use to place probability / make decisions in regions of input space with little data?

Interpretability: once we have a trained network, how do we understand what it does? How is the training data embedded in the weights?

Biological Plausibility: how can we do what we do within the constraints of neurobiology? How can we interpret specific architectures used by the brain?

References

- Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, Cog Sci. 2013.
- Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, ICLR 2014.
- Identifying and attacking the saddle point problem in high dimensional non-convex optimization, Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio, NIPS 2014.
- A memory frontier for complex synapses, S. Lahiri and S. Ganguli, NIPS 2013.
- M. Advani and S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions, Physical Review X 2016.
- Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, under review, NIPS 2016.
- Deep information propagation, S. Schoenholz, J. Gilmer, S. Ganguli, J. Sohl-Dickstein, ICLR 2017.
- On the expressive power of deep neural networks, M. Raghu, B. Poole, J. Kleinberg, J. Sohl-Dickstein, S. Ganguli, ICML 2017.
- Continual learning through synaptic intelligence, F. Zenke, B. Poole, and S. Ganguli, ICML 2017.

• <http://ganguli-gang.stanford.edu>

Twitter: @SuryaGanguli

The project that really keeps me up at night

